2007

Examining the issues surrounding violating the assumption of independent observations in reliability generalization studies: A simulation study

Jeanine L. Romano
University of South Florida

Examining the Issues Surrounding Violating the Assumption of Independent Observations in Reliability Generalization Studies: A Simulation Study

by

Jeanine L. Romano

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Measurement and Evaluation
College of Education
University of South Florida

Major Professor: Jeffrey D. Kromrey, Ph.D.
Michael T. Brannick, Ph.D.
Robert F. Dedrick, Ph.D.
Anthony J. Onwuegbuzie, Ph.D.

Date of Approval
February 26, 2007

Keywords: Vacha-Haase, Research Synthesis, Meta-analysis, Intra-class Correlation, Statistical Assumptions

Dedication

This dissertation is dedicated to the memory of my mother June Alice Romano, the smartest person I have ever known. She was a woman who raised six children and loved us all unconditionally. She lived her live with passion. I aspire every day to be just like her. My mother taught me that the key to success is perseverance. There were plenty times during this journey that I had my doubts about my ability to complete my PhD. During the times where I felt great self doubt and somewhat overwhelmed, I would find comfort in remembering what my mother always said to me when I was so very young: “Anything is possible if you put your mind to it”! For the graduate student that might be perusing this document looking for encouragement for their dissertation, I hope her words will inspire you as well!

Acknowledgements

This dissertation would not have been possible if it hadn't been for the many individuals that have contributed to my learning along the way. First, I want to thank my brothers and sisters: Geno, Joanne, George, Vincent and Patty, they have been there to encourage me along the way. Second, I would also like to thank my boss and dear friend, Dr. Linda Devine, who has supported me both professionally and personally. In addition, she has been the ultimate role model for women in the field of Higher Education; I want to be just like her when I grow up! Third, I would like to thank my friends Lisa, David, and Janene that allowed me to invade their computers with my strange SAS software in order to run some of the conditions for this study. Fourth, I'd like to thank my USF "family" my fellow graduate students: Moya, Melinda, Gianna, Kris, Peggy, Tom, Susan, Bruce, Ha, Dorian, and Jesse; our department office manager: Lisa Adkins; and the other department faculty members: Dr. Bruce Hall, Dr. Lou Carey, Dr. Connie Hines, and Dr. John Ferron. They have contributed to my learning in ways that I can not even begin to measure!

Finally, I'd like to thank the members of my committee. Michael Brannick for showing me that research should be exciting. Robert Dedrick for showing me that research should be comforting. Anthony J. Onwuegbuzie for showing me that research should be precise. Finally, I'd like to thank Jeffrey D. Kromrey, my mentor and major professor. Even when the early drafts of this document where barely comprehendible, he would always say, "It's a good start." I will always be grateful for his encouraging words and his unlimited patience! He taught me so much about research and even more about being a kind and patient human being. It has truly been an honor having him as my major professor and colleague.

TABLE OF CONTENTS

List of Tables iii

List of Figures vi

Abstract viii

Chapter One: Introduction 1

Methodological Issues in RG Studies 2
Purpose of the Study 4
Research Questions 5
Hypotheses 6
Limitations of the Study 8
Definitions of Terms 9
Importance of the Study 12

Chapter Two: Literature Review 13

Meta-Analysis 13
The Reliability of Measures 18
RG Studies 21
Issues in the Debate on Reliability Generalization 34
Independence 39
Summary 45

Chapter Three: Method 48

Purpose...48
Research Questions 49
Sample..49
Method 52
Simulation of the data. 53
Simulation of intra-class correlation. 55
Conduct of RG analyses..58
Evaluation of the results..59

Chapter Four: Results 62

How the Results were Evaluated 62
Box Plots...64
Summary of Box Plot Results...68
η^2 Analysis 69
Bias 71

Root Mean Squared Error 75
Band Coverage 80
Band Width 89
A Deeper Look at Band Coverage 96
Summary of Results 102

Chapter Five: Conclusions 108

Summary of the Study 108
Research Questions 111
Summary of Study Results 112
Limitations of the Study 118
Implications 119
Importance of the Study. 119
Importance in Regards to Future RG Studies. 119
Suggestions for Future Research 121

References 126

Appendices: 143

Appendix A: SAS Code for Monte Carlo Simulation 144

About the Author .... End Page

List of Tables

Table 1 Reliability Generalization Studies 22

Table 2 Number of Items Needed to Generate Reliability Parameter 50

Table 3 Population Item Parameters Used for Simulations 51

Table 4 Results of the Simulation of Intra-class Correlation..55

Table 5 Correlation Between Dependent Variables..63

Table 6 η^2 Analysis of the Effects of Factors in the RG Simulation.................................70

Table 7 Bias in Estimated Mean Reliability by Treatment and Intra-class Correlation 71

Table 8 Bias in Estimated Mean Reliability by Treatment and Coefficient Alpha..72

Table 9 Bias in Estimated Mean Reliability by Treatment and Average Sample Size..73

Table 10 Bias in Estimated Mean Reliability by Intra-class Correlation and Coefficient Alpha...74

Table 11 RMSE of Estimated Mean Reliability by Treatment and Coefficient Alpha..76

Table 12 RMSE of Estimated Mean Reliability by Treatment and Average Sample Size..76

Table 13 RMSE for Estimated Mean Reliability by Treatment and Intra-class Correlation 77

Table 14 RMSE for Estimated Mean Reliability Average Sample Size by Coefficient Alpha.....78

Table 15 RMSE for Estimated Mean Reliability Intra-class Correlation by Coefficient Alpha.....79

Table 16 Band Coverage of Estimated Mean Reliability by Treatment and Intra-class Correlation.....81

Table 17 Band Coverage of Estimated Mean Reliability for Intra-class Correlation and Treatment by Coefficient Alpha.....83

Table 18 Band Coverage of Estimated Mean Reliability for Average Sample Size and Treatment by Coefficient Alpha.....86

Table 19 Band Coverage of Estimated Mean Reliability for Number of Primary Studies by Treatment.....88

Table 20 Band Coverage of Estimated Mean Reliability for Coefficient Alpha by Treatment.....88

Table 21 Band Width of Estimated Mean Reliability by Treatment and Average Sample Size.....90

Table 22 Band Width of Estimated Mean Reliability by Treatment and Coefficient Alpha.....90

Table 23 Band Width of Estimated Mean Reliability by Treatment and Number of Primary Studies.....91

Table 24 Band Width of Estimated Mean Reliability for Average Sample Size by Coefficient Alpha.....92

Table 25 Band Width of Estimated Mean Reliability for Number of Primary Studies by Average Sample Size 93

Table 26 Band Width of Estimated Mean Reliability for Number of Primary Studies by Coefficient Alpha 95

Table 27 Large Band Coverage by Type of Treatment 97

Table 28 Percentage of Large Band Coverage for Intra-class Correlation by Treatment 98

Table 29 Percentage of Large Band Coverage for Coefficient Alpha by Treatment 99

Table 30 Percentage of Large Band Coverage for Average Sample Size by Treatment 100

Table 31 Percentage of Large Band Coverage for the Number of Primary Studies by Treatment 102

Table 32 Percentage of Large Band Coverage for Number of Reliabilities Per Study by Treatment 102

List of Figures

Figure 1. The model of the simulation for the study..54

Figure 2. A model for how the “test-taking” was simulated...57

Figure 3. Distribution of bias estimates for reliability coefficients for all five types of treatments..65

Figure 4. Distribution of RMSE estimates for reliability coefficients for all five types of treatments 66

Figure 5. Distribution of band coverage for reliability coefficients for all five types of treatments 67

Figure 6. Distribution of band width for reliability coefficients for all five types of treatments 68

Figure 7. Bias in estimated mean reliability by treatment and coefficient alpha...72

Figure 8. Bias in estimated mean reliability by intra-class correlation and coefficient 74

Figure 9. RMSE for estimated mean reliability average sample size by coefficient alpha...78

Figure 10. RMSE for estimated mean reliability intra-class correlation by oefficient alpha..79

Figure 11. Band coverage of estimated mean reliability intra-class orrelation by coefficient alpha 84

Figure 12. Band coverage of estimated mean reliability for average sample size by coefficient alpha...87

Figure 13. Band width of estimated mean reliability for average sample size by coefficient alpha...92

Figure 14. Band width of estimated mean reliability for number of primary studies by average sample size..94

Figure 15. Band width of estimated mean reliability for number of primary studies by coefficient alpha...95

Examining the Issues Surrounding Violating the Assumption of Independent
Observations in Reliability Generalization Studies:
A Simulation Study

Jeanine L. Romano

ABSTRACT

Because both validity and reliability indices are a function of the scores on a given administration of a measure, their values can often vary across samples. It is a common mistake to say that a test is reliable when in fact it is not the test that is reliable but the scores on the test that are reliable. In 1998, Vacha-Haase proposed a fixed-effects meta-analytic method for evaluating reliability that is similar to validity generalization studies called reliability generalization (RG). This study was conducted to evaluate alternative analysis strategies for the meta-analysis method of reliability generalization when the reliability estimates are not statistically independent. Five approaches for handling the violation of independence were implemented: ignoring the violation and treating each observation as independent, calculating one mean or median from each study, randomly selecting only one observation per study, or using a mixed effects model. This Monte Carlo study included five factors in the method. These factors were (a) the coefficient alpha, (b) sample size in the primary studies, (c) number of primary studies in the RG study, (d) number of reliability estimates from each, and (e) the degree of violation of independence where the strength of the dependence is related to the number of reliability indices (i.e. coefficient alpha) derived from a simulated set of examines and the magnitude of the correlation between the journal studies (with intra-class correlation

ICC = 0, .0l , .30, and .90). These factors were used to simulate samples under known and controlled population conditions. In general, the results suggested that the type of treatment does not have a noticeable impact on the accuracy of the reliability results but that researchers should be cautious when the intra-class correlation is relatively large. In addition, the simulations in this study resulted in very poor confidence band coverage. This research suggested that RG meta-analysis methods are appropriate for describing the overall average reliability of a measure or construct but the RG researcher should be careful in regards to the construction of confidence intervals.

Chapter One:

Introduction

Ideally, social science research is conducted using measurement instruments that will produce valid and reliable information. When these tests are first developed to measure a certain construct (e.g., depression), they are usually evaluated in regard to the validity and reliability of their scores. Although these procedures are conducted for the development of the instrument, the fact that both validity and reliability can fluctuate across samples, as both indices are a function of the scores on a given administration of a measure, is often overlooked. It is a common mistake to report that a test is reliable when in fact it is not the test that is reliable but the scores on a test that are reliable (Vacha-Haase, Kogan, & Thompson, 2000).

Because reliability can fluctuate across studies, it has been recommended that researchers should always evaluate the reliability of their scores and report the results. The American Psychology Association (APA) Task Force on Statistical Inference in their 1999 report stated:

> It is important to remember that a test is not reliable or unreliable. Reliability is a Property of the scores on a test for a particular population of examinees…Thus, authors should provide reliability coefficients of the scores for the data being analyzed even when the focus of their research is not psychometric (Wilkinson & APA Task Force on Statistical Inference, 1999, p. 596).

Validity generalization studies have been conducted to describe the extent to which validity evidence for scores are generalizable across research contexts (Hunter & Schmidt, 1990; Schmidt & Hunter, 1977). In 1998, Vacha-Haase proposed a fixed-effects meta-analytic method for evaluating reliability, similar to validity generalization studies, this is called reliability generalization (RG). RG studies can be used to investigate the distribution of reliability estimates across studies and to identify study characteristics that may be related to variation in reliability estimates, such as sample size, type of reliability estimate (coefficient alpha vs. test-retest), different forms of an instrument, or participant characteristics (Henson, 2001; Vacha-Haase, 1998). This method is recommended for describing estimated measurement error in a test's scores across studies and can also be used to analyze measurement error in different scales that measure the same construct.

Methodological Issues in RG Studies

Potential methodological problems are evident in RG studies, and the debate about their solution has only just begun (Helms, 1999; Sawilowsky, 2000; Thompson & Vacha-Haase, 2000). Major controversies include (a) approaches for treatment of large proportions of missing data in the published literature, (b) the use of nonlinear transformations of sample reliability estimates, (c) the need to weight the observed sample statistics to account for differences in sampling error across studies (d) the differences between analyses of reliability coefficients and analyses of the estimated standard errors of measurement (SEM), and (e) appropriate analyses of reliability estimates that are not statistically independent (Sawilowsky, 2000; Thompson & Vacha-Haase, 2000).

This research primarily focused on appropriate analysis of reliability estimates that are not statistically independent. Several RG studies have been conducted that included samples that did not represent independent observations. For example, in their study on the State-Trait Anxiety Inventory (STAI; Spielberger, Gorsuch, & Lushene, 1970), Barnes, Harp, and Jung (2002) obtained 117 reliability coefficients from 45 articles where each subgroup of participants was treated as an observation. When Capraro and Capraro (2002) conducted an RG study on the Myers-Briggs Type Indicator scale (Myers & McCaulley, 1985), they included 70 reliability coefficients from only 14 published studies. Yin and Fan's (2000) RG study on the Beck Depression Inventory (BDI; Beck, Ward, Mendelson, Mock, & Erbaugh, 1961) included 164 reliability coefficients from 90 studies. Similarly, Vacha-Haase's (1998) RG study on the Bem Sex Role Inventory (BSRI; Bern, 1974) used 87 reliability coefficients from 57 studies, and Caruso's (2000) RG study on the NEO personality scale (Costa & McCrae, 1985) used 51 reliability estimates from 37 studies. Clearly, these are violations of independence of observations.

The assumption of independence of observations is commonly violated in meta-analytic research (Becker, 2000; Hedges & Olkin, 1985; Hunter & Schmidt, 1990). Violations can lead to biased results in which Type I error (rejecting a true null hypotheses) and Type II error (failing to reject a false null hypotheses) rates are inaccurate (Barcikowski, 1981; Scariano & Davenport, 1987). The problem of violation of independence has been investigated in regard to various statistical techniques (e.g., Barcikowski, 1981; Bock, 1975; Hewitt-Gervais & Kromrey, 1999; Kenny & Judd, 1986; Kromrey & Dickinson, 1996; Raudenbush & Bryk, 1987; Scariano & Davenport, 1987).

Further, several studies have been conducted concerning the consequences of dependent observations in meta-analysis (e.g., Becker & Kim, 2002; Beretvas & Pastor, 2003; Cooper, 1979; Greenhouse & Iyengar, 1994; Hedges & Olkin, 1985; Landman & Dawes, 1982; Raudenbush, Becker, & Kalaian, 1988; Rosenthal & Rubin, 1986; Tracz, Elmore, & Pohlmann, 1992). In general, this body of research has indicated that ignoring the assumption of independence can impact the magnitude of statistical significance.

There are several approaches to dealing with the violation of independence that have been recommended by researchers (Becker, 2000). These approaches include, ignoring it and treating each observation as independent (e.g., Smith, Glass, & Miller, 1980), calculating one mean or median from each study (e.g., Tracz et al., 1992), selecting only one observation per study (e.g., Rosenthal & Rubin, 1986), and using a mixed effects model (e.g., Beretvas & Pastor, 2003).

As the available literature suggests, violating the assumption of independence is a serious issue. Because the RG study method is a relatively new technique, it is imperative that the consequences of violating independence be investigated. Even more important, the research techniques that have been used in previous treatments to control for violation of independence need to be investigated in the RG study environment to investigate the extent to which Type I error is impacted.

Purpose of the Study

This study's purpose was to examine the potential impact of selected methodological factors on the validity of RG study conclusions. Although all of the controversies described previously are important, this study focused on the issues surrounding violating the assumption that the observations are independent and the methods that researchers have devised to handle dependent data in a meta-analysis. Factors such as (a) the magnitude of coefficient alpha, (b) sample size (i.e., number of examinees), (c) number of studies, (d) the number of reliabilities included in each journal study, and (e) the magnitude of the intra-class correlation among journal studies (i.e. the degree of dependence among journal studies) were also considered. The impact of these factors on the accuracy of estimating reliability was investigated when four approaches to violation of independence were used: (a) treating dependent observations as independent, (b) randomly selecting a reliability index from each study, (c) calculating a mean or a median, and (d) using a two-level mixed effects model. In other words, for certain method factors, does violation of independence significantly impact the accuracy of estimating the true reliability parameter?

Research Questions

In RG studies, the dependent variable in the analyses is the reliability estimate (Henson & Thompson, 2001). This research focused on how certain study methods, in regards to violation of independence, affect the estimated mean reliability of scores calculated across studies. The key questions that were addressed in this study were:

1. What is the effect on point and interval estimates of mean reliability of ignoring violation of independence of observations in RG studies (i.e., treating all reliability coefficients as independent observations)?
2. What is the effect on point and interval estimates of mean reliability of using a mean or median reliability from each study as part of a sample in a RG study?
3. What is the effect on point and interval estimates of mean reliability of randomly selecting a reliability estimate from each study as a part of a sample in a RG study?

4. What is the effect on point and interval estimates of mean reliability of using a two level mixed-effects model for RG studies (i.e., reliabilities are nested within studies)?
5. In regard to violations of independence, what impact do factors such as the magnitude of coefficient alpha, sample size, number of journal studies, number of reliability coefficients from each study, and the magnitude of the intra-class correlation (ICC) of the studies (i.e., the magnitude of the violation of independence) have when any of the methods discussed in the four research questions above are investigated?

Hypotheses

1. Of the five approaches to dealing with violation of independence examined in this research, ignoring the dependence among studies provides the worst point and interval estimates of the reliability in RG meta-analysis compared to the other treatments used; confidence interval coverage will be grossly underestimated when dependence is ignored.
2. Randomly selecting one reliability estimate from each study as a means to control for dependence provides better point and interval estimates of the reliability in the RG meta-analysis than ignoring the dependence; confidence interval coverage will be less problematic when randomly selecting one reliability coefficient from each study than when the dependence is ignored.
3. Calculating a mean or a median reliability from each study as a means to control for dependence provides better point and interval estimates of the reliability in the RG meta-analysis than randomly selecting one reliability estimate from each study and even better point and interval estimates than ignoring the violation of independence;

confidence interval coverage will be less problematic using this method than when using the other previous methods (ignoring, randomly selecting).

4. The use of a two-level mixed model provides better point and interval estimates of the reliability than the other four approaches examined in this research; the two-level mixed model is the best approach for confidence interval coverage in regards to violation of independence in RG meta-analysis.
5. While ignoring the dependence is the worst approach and the use of the two-level mixed model is the best approach for estimating point and interval estimates of the reliability, the extent to which the above methods are tenable will be moderated by the following characteristics in the RG meta-analysis.
   a. Point and interval estimates generated from population with larger reliability coefficients are less biased than are those estimates generated from populations with smaller reliability coefficients; as the reliability estimate increases the bias of the point and interval estimates decreases.
   b. Point and interval estimates generated from populations where the mean sample size of groups is small are more biased than are those estimates generated from populations where the mean sample size is large; as the mean sample size of groups increases the bias of the point and interval estimates decreases.
   c. Point and interval estimates generated from populations where the number of journal studies is large are less biased than those generated from populations where the number of journal studies is small; as the number of journal studies increases the bias of the point and interval estimates decreases.

d. Point and interval estimates generated from populations where the number of reliabilities is large are more biased than those generated from populations where the number of reliabilities is small; as the number of reliabilities from each study increases the biased of the point and interval estimates increases.

e. Point and interval estimates generated from populations where the intra-class correlation is large are more biased than are those estimates from populations where the intra-class correlation is small or zero; as the intra-class correlation increase the bias of the point and interval estimates increases.

Limitations of the Study

The limitations of this study are related to the Monte Carlo method for the study. While the Monte Carlo method was used to simulate RG studies, the values of the factors used in the simulation were fixed for each study. Because the data for this study were simulated, the number of reliability indices from each simulated study was a fixed value in each of the simulations as each study contributed the same number of reliability indices per study. While it is obvious that several of the RG studies conducted previously treated reliability coefficients from the same study as independent, not all of the studies contributed equal numbers of reliability coefficients.

In several of the RG studies conducted previously, test-retest reliability estimates given are very rarely and seldom evaluated. Because coefficient alpha is the most common reliability coefficient reported, this was the only index used in the study. It is important to remember, however, that coefficient alpha has a tendency to under estimate the actual reliability index (Crocker & Algina, 1986).

Definitions of Terms

The following terms are used throughout this study:

Classical Test Theory - A model used in testing where an individual's observed score (X) on a measure is composed of the sum of his or her true score (T) and error score (E), $\left(\text{i.e.}\,, X = T + E;\ \text{Crocker \& Algina,1986}\right)$.

Effect Size - The magnitude of the effect of a treatment. According to Cohen (1988) it is "the degree to which a phenomenon is present in a population" (p. 78).

Types of Effect Size - There are many ways to calculate an effect size. However, Rosenthal (1994) states they basically fall into two "families": the *d* family and the *r* family. The *d* family is based on Cohen's *d* where *d* is the sample effect size that estimate the population effect size *,* δ such that $\delta = \dfrac{\mu_E - \mu_C}{\sigma_C}$ and $d = \dfrac{\overline{X}_E - \overline{X}_C}{s_C}$. The *r* family refers to the Pearson product moment correlation.

Intra- Class Correlation (ICC) - The statistical index that measures the magnitude of the dependence among observations such that: $ICC = \dfrac{(MS_b - MS_w)}{(MS_b - (i-1)MS_w)}$ where MS$_b$ is the mean squares between studies, MS$_w$ is the mean squared within studies, and *i* is the number of reliabilities for each study (Stevens, 1999). This value can range from $-\dfrac{1}{n-1}$ to 1. The larger the ICC the higher the degree of the dependence (Kenny & Judd, 1986). In mixed models, where there is a two-level hierarchy, it is defined as the proportion of variance in the dependent variable that is between the second-level units (Kreft & de Leeuw, 1998). Specifically, in an RG study level one would model the variance within studies and level two would model the variance between studies. Therefore, the ICC would represent the proportion of variance in reliability that is between studies.

Independence of Observations - The assumption that observations in a study are independent means there is no correlation or relationship between them (Glass & Hopkins, 1996). Kenny and Judd (1986) define independence of observations in term of conditional probabilities: “If two observations are independent of each other, then the conditional probability of one of them, given the other, is not different from the unconditional probability” (p. 422). If X_i and X_j are samples from an infinite population with a mean of μ and a variance of σ^2, then the observations X_i and X_j are said to be independent if the expected value of the product of the distance of X_i to the mean and the distance of X_j to the mean is equal to zero $\left(\text{i.e. } E\left[(X_i - \mu)(X_j - \mu)\right] = 0\right)$ (Kenny & Judd, 1986). In the case of meta-analytic research, observations are considered to be independent when the value of any statistic (when it is included in a meta-analysis) is in no way predictable from the value of any other included statistic in the same meta-analysis study (Tracz et al., 1992).

Meta-Analysis -This is the method developed by Glass (1976) that uses statistical procedures to combine results of multiple studies to make inferences in regards to an overall measure of an index (e.g., reliability) across studies.

Mixed Effects Models - This is a model that is used for multiple levels of measurements that analyzes data in a clustered or nested structure (Kreft & de Leeuw, 1998). It is often referred to as Hierarchical Linear Modeling (HLM).

Reliability -This is the degree to which the scores of a measure (i.e. test) are consistent over repeated administrations of the same test or parallel forms of the test (Crocker & Algina, 1986).

Reliability Index - This is the correlation that represents the strength of the relationship between true and observed scores. It is the ratio of the standard deviation of true scores to the standard deviation of observed scores $\left(\rho_{\mathrm{XT}} = \frac{\sigma_T}{\sigma_x} \right)$ (Crocker & Algina, 1986).

Reliability Coefficient - This is the ratio of true score variance to observed score variance and is the square of the reliability index. It is also defined as the correlation between two scores on parallel tests $\left(\rho_{\mathrm{X}_1 X_2} = \frac{\sigma_T^2}{\sigma_X^2} \right)$ (Crocker & Algina, 1986). There are two main types of reliability coefficients:

> Test-Retest - This is the correlation between scores on two separate administrations of the same measure given to same group of individuals.
>
> Internal-Consistency - This is the correlation that is based on a single administration of a test.

The types of reliability coefficients will be discussed in detail in chapter 2.

Reliability Generalization (RG) Studies - This is a meta-analysis study method that was developed by Vacha-Haase to make generalizations about the average reliability of a measure or construct (Vacha-Haase, 1998).

Reliability Induction - When authors report reliability from previous samples or test manuals when defending the reliability of the data in their studies (Vacha-Haase, 1998).

Importance of the Study

Whereas Thompson and Vacha-Haase (2000) have argued that a series of RG studies could reveal that, across samples, the reliability of scores for a given scale are relatively stable, they also reported that it is possible that such analyses could reveal that the variation in reliability is not related to treatment factors. It is important to recognize that to comprehend what RG studies may reveal, the consequences of the method flaws of RG studies must first be considered. This research will address the consequences of violating the assumption of independence and offer some suggestions for handling these issues. Not only will the results of this research contribute to the future RG study methods, it will also serve as a reminder of the consequences of ignoring important assumptions in all research.

Chapter Two:

Literature Review

This literature review is divided into four parts. First, meta-analysis and reliability are briefly discussed. Second, literature on specific points of interest on the 32 published RG studies that have been conducted to date is presented. Third, the issues that have been addressed by scholars in response to RG studies method are presented. Finally, the literature about violation of independence is presented.

Meta-Analysis

Meta-analysis, sometimes referred to as research synthesis, is a quantitative research approach that converts individual study outcomes to a common metric, such as effect sizes, and compares them across studies. Each study is considered one observation from a hypothetical universe of studies. In 1976, Glass originated the term 'meta-analysis' and defined it as "the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating findings" (p. 3).

Meta-analysis is a secondary analysis that can be used to summarize quantitatively large bodies of literature. When a large number of studies are aggregated, meta-analysis can investigate factors that were not investigated in the primary studies and detect the effect of possible moderating variables. Since it was first introduced, several approaches have been developed. There are five basic approaches: classical or Glassian meta-analysis, study effects meta-analysis, homogeneity test-based meta-analysis, validity generalization meta-analysis, and psychometric meta-analysis (Bangert-Drowns, 1986; Hunter & Schmidt, 1990).

The classical or Glassian meta-analysis procedure calculates the mean effect size, ($\bar{d}$*),* as an estimate of the population effect size, (δ), across studies for the entire universe of studies. This method has been criticized for being too liberal when determining which studies to include in a meta-analysis. Glass argued that all the studies related to a given topic should be in the sample regardless of quality of a study because study quality is related to the variance of treatment effects in each study. In Glass's method, the unit of analysis is the study finding such that effect sizes can be calculated for each comparison between groups or sub groups for the different criteria from each individual study. In Glass's method, effect sizes also can be averaged from different dependent variables that may measure different constructs (Bangert-Drowns, 1986; Hunter & Schmidt, 1990).

The Glassian method has been strongly criticized for several reasons: (a) it clearly violates the assumption of independence of observations by including several effect sizes from a single research study, which, in turn, leads to rather large inflated total sample size; (b) all studies are included regardless of the quality of the method (i.e., "garbage in garbage out"); and (c) the method has a tendency to mix different independent and dependent variables (Bangert-Drowns, 1986; Hedges, 1982; Hunter & Schmidt, 1990). Glass defended his methodology by stating the purpose of a meta-analysis is to present a very broad overview of a specific research interest. For example, Glass investigated the impact of all types of psychotherapy (Smith, Glass, & Miller, 1977) and the effects of class size on all types of achievement (Glass & Smith, 1979).

The study effect meta-analysis is very similar to the Glassian method except the criteria for the inclusion of studies are much more selective. If a study's methods are flawed, the study is not included. Another important difference is the study is the unit of analysis; thus, only one effect size is calculated for each study. This method was suggested by Mansfield and Busse (1977) and has been applied in several meta-analyses since 1979 (see Bangert & Drowns, 1986).

Some researchers advocate inclusion rules that are more selective (Henson, Kogan, & Vacha-Haase, 2001; Landman & Dawes, 1982; Mansfield & Busse, 1977; Thompson & Vacha-Haase, 2000; Vacha-Haase, 1998; Wortman & Bryant, 1984). Studies with serious methodological flaws are excluded. The difficulty with this approach is that reviewer bias can influence decisions about which studies should be included. This, in turn, may distort the findings of the meta-analysis in regard to the true population. Glass argued that all the studies related to a given topic should be in the sample, regardless of quality. The distribution of effect sizes should then be corrected for sampling error, measurement error, and restriction of range. Study reports or publications, however, do not always contain information necessary for making these corrections. Meta-analysts using this approach may average effects from different dependent variables, even when these effects measure different constructs. The problem is that when study findings are used as the units of analysis, non-independent data are produced and greater weight is given to studies with more comparisons. This may cause a bias towards statistically significant results (i.e., inflation of Type I error rate; Bangert-Drowns, 1986).

The test of homogeneity meta-analysis method is used to evaluate how much of the variance among effect sizes is due to sampling error. In this method, statistical tests are used to determine if the variability in study outcomes is statistically significant. If the tests are statistically significant, then this would be the basis for detecting moderating variables (Bangert-Drowns, 1986; Hedges, 1982; Hunter & Schmidt, 1990; Rosenthal & Rubin, 1982). A major criticism of this approach is that it based on only the estimated error and lacks the power to detect differences. Hunter and Schmidt (1990) contended that there could be other artifacts that might be sources of variance. Hedges and Olkin (1985) contended that even if the variance across studies is statistically significant and not due to artifactual sources, it is often small in magnitude and usually not practically significant. They cautioned researchers to investigate the actual size of the variance.

Schmidt and Hunter (1977) developed a procedure usually referred to as validity generalization to address the problem of artifacts that can affect variance in observed effect sizes. In this particular method, correlations are used to measure effect sizes. Schmidt and Hunter argued that the mean effect size should be corrected because it was attenuated by unreliability and possible range restriction. In this method, they test for statistical artifacts. There are 11 statistical artifacts which Hunter and Schmidt (1990) have identified that could distort the size of the study correlation. These are (a) sampling error, (b) error of measurement in the dependent variable, (c) error of measurement in the independent variable, (d) dichotomization of a continuous dependent variable, (e) dichotomization of a continuous independent variable, (f) range variation in the independent variable, (g) range variation in the dependent variable, (h) deviation from perfect construct validity in the independent variable, (i) deviation from perfect construct validity in the dependent variable, (j) reporting or transcription error, and (k) variance due to extraneous factors. Hunter and Schmidt (1990) argue that if the first three of these artifacts account for 75% or more of the observed variance of the effect sizes, the residual of the observed variance is due to the other eight. This led to conclusion that the true observed variance was actually zero (i.e., the effect sizes from each study are homogeneous). If this is not the case, the next step should involve testing for moderating variables.

RG studies involve looking at measurement error across studies in an attempt to characterize the psychometric properties of the hypothetical universe of studies that may employ a particular measure. Such properties may include the mean reliability coefficient obtained in such a population, the variance of the reliability coefficient across studies, and treatment factors that may influence the magnitude of the coefficient (i.e., moderating variables).

In the aggregation of research results through meta-analysis, fundamental questions typically focus on (a) point and interval estimation of the mean effect size and (b) the relationship between the mean effect size and treatment factors. Estimates of mean effect sizes and relationships between effect sizes and other variables usually are obtained using weighted least squares, in which effect sizes of individual studies are weighted by the inverse of their sampling variance (Hedges & Olkin, 1985). That is,

$$v_i = \frac{1}{\mathrm{var}\left(\hat{\delta}_i\right)}$$

where v_i = weight for the *i* effect size, and

$\mathrm{var}\left(\hat{\delta}_i\right)$ = estimated sampling variance of the *i* effect size.

The argument for using such weights in statistical estimates is that the weights will give greater credibility to the effect sizes obtained from studies with less sampling error. These studies typically have larger sample sizes.

Glass argued that literature reviews should be as systematic as primary research and should interpret the results of individual studies in the context of distributions of findings, partially determined by study characteristics and partially random. Since that time, meta-analysis has become a widely accepted research tool encompassing a family of procedures used in a variety of disciplines. In a meta-analysis, research studies are collected, coded, and interpreted using statistical methods similar to those used in primary data analysis. The result is an integrated review of findings that is more systematic and exact than a narrative review.

The Reliability of Measures

Reliability refers to dependability or consistency. In educational and psychological research, tests are used to quantify the relative standing of an individual on a psychological trait or ability. In educational and psychological research when attempts are made to measure a trait or ability more than once for an individual, it is very unusual for that individual to score exactly the same for every administration, unlike the physical sciences. What can be measured is the degree to which a test score is consistent. The challenge is that when individuals take a test, there are systematic and random errors that can occur when a test is repeated. In addition, repeated administrations of a test are not always feasible. In classical test theory, the reliability coefficient, ρ_{xx}, is defined as the correlation between scores on parallel tests (Crocker & Aligina, 1986). According to classical test theory, an examinee's observed score, X, can be expressed as the sum of his/her true score and random error:

$$X = T + E$$

The reliability coefficient is the proportion of the observed variance in scores that represents true score variance rather than random error:

$$\rho_{xx} = \frac{\sigma_{true}^2}{\sigma_{total}^2}$$

where ρ_{xx} is the ratio of the true score variance to total score variance.

The most common approaches for estimating the reliability of scores include administering the same test twice to the same examinees (test-retest reliability) or administering the test once and estimating score reliability from the intercorrelation of test items (internal consistency reliability). Test-retest reliability is estimated by calculating the correlation coefficient between the scores obtained on the two administrations of the test. Internal consistency reliability is estimated by calculating the correlations between subsets of items on the test (Crocker & Algina, 1986). There are several indices that can be used to measure internal consistency:

Coefficient Alpha - Also known as Cronbach's alpha, it can be calculated as follows: $\alpha = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum s_i^2}{s_x^2}\right)$ where *k* is the number of items on a test and s_i^2 is the variance of item *i,* and s_x^2 is the total test variance (Crocker & Algina, 1986).

Kuder Richardson Formulas (KR21 and KR20) - These are indices of homogeneity that Kuder and Richardson (1937) developed that are based on the proportion of correct and incorrect answers to each of the items on the test. Kuder Richardson Formulas are used when a test is scored dichotomously. KR20 can be calculated as follows:

$$\text{KR20} = \left(\frac{k}{k-1}\right)\left(1-\frac{\sum pq}{s_x^2}\right),$$

where k is the number of items on the test s_x^2 is the variance of scores on the total test, p is the proportion of correct answers, and q is the proportion of incorrect answers. KR 21 is similar to KR20 except with KR21 it is assumed that all items on a measure are equally difficult. KR21 can be calculated as follows:

$$\text{KR21} = \left(\frac{k}{k-1}\right)\left(1-\frac{\bar{X}(k-\bar{X})}{ks_x^2}\right),$$

where k is the number of items on the test, s_x^2 is the variance of scores on the total test, and $\bar{X}$ is the mean of the scores.

Split-half Method - Reliability is estimated by artificially splitting a measurement in half and calculating the correlation between the two halves. It has been argued that this produces a reliability coefficient that underestimates the true reliability (Crocker & Algina, 1986), therefore the Spearman Brown prophecy formula can be employed to calculate a corrected estimate. It can be calculated as follows:

$$\rho_{xx'} = \frac{2\rho_{AB}}{1+\rho_{AB}},$$

where $\rho_{xx'}$ is the predicted reliability coefficient for the full-length of the test and ρ_{AB} is the correlation between the two halves.

RG Studies

Since 1998, 32 RG studies have been published; these have been labeled with an asterisk in the reference section and are listed in Table 1. In addition the scales in which these RG studies have examined are labeled with a double asterisk in the reference section. Of these, only three (Henson et al., 2001; Reese, Kieffer, & Briggs, 2002; Viswesvaran & Ones, 2000) have examined reliability generalization in terms of multiple measures of the same construct. The following paragraphs highlight some of the key characteristics of these 32 studies.

Inclusion criteria are key characteristics of RG studies. In most cases, before a study was examined in terms of its reliability reporting, studies had to be in English and published. There were only three RG studies that allowed non-English studies into the sample (Barnes, Harp, & Jung 2002; Beretvas, Meyers, & Leite, 2002; De Ayala, Vonderharr-Carlson, & Kim, 2005) and only five articles that included dissertations in the sample (Barnes et al., 2002; Beretvas et al., 2002; Capraro & Capraro, 2002; Nilsson, Schmidt, & Meek, 2002; O'Rourke, 2004). Very little explanation was given in any of the RG studies to support the inclusion criteria. When Caruso and Edwards (2001); Caruso, Witkeiwitz, Belcourt-Dittloff and Gottlieb (2001) and Leach, Henson, Odom, and Cagle (2006) conducted their RG studies not only did they use only published studies in English, they also eliminated any test-retest coefficients. It was not clear in these articles if the test-retest coefficients came from the same articles as the alphas in the RG study or if they were from other published studies. The authors simply argued that there were not enough test-retest coefficients to conduct a valid study.

Table 1

Reliability Generalization Studies

Study and Instrument (or construct) investigated	**Additional inclusion criteria**	**Number of articles found that reported reliability for the data "in hand"**	**Number of samples and type of reliability**	**Reliability Mean, Median and Standard Deviation.**	**Type of Analysis**
Barnes, Harp, & Jung (2002) Spielberger State-Trait Anxiety Inventory	Published articles only Allowed non-English articles	46 articles 6 % of the articles found	Total number of samples 117. 59 were state (52 alpha, 7 test-retest) 58 were trait (51 alpha, 7 test-retest)	State Alpha M= .91; Md= .92; *SD*= .05 Test-retest M= .70; Md = .68; *SD* = 0.20 Trait Alpha M=.89; Md = .90; *SD*= .05 Test-retest M= .88; Md = .88; *SD* = 0.05	Descriptive statistics for both alpha and test-retest presented separately Correlation for alpha only
Beretvas, Meyers, & Leite (2002) Marlowe-Crowne Social Desirability Scale	Published articles and dissertations Allowed non-English articles	72 articles 8.7% of the studies found	Total number of sample 182 149 Cronbach's alpha 3 Spearman Brown 9 KR20 21 test-retest	Mixed effects model M= .726; SE = .0248 Fixed effects model M= .68; SE = .0059 Median not reported	Fisher-z transformation applied Mixed effect models Internal consistencies grouped together Only 123 internal consistency reliabilities were used in the mixed effect model
Capraro & Capraro (2002) Myers-Briggs Type Indicator	Published articles and dissertations No mention of non-English versions	14 articles 7% of the articles found	Total number of samples 70 50 alpha 20 test-retest	Alpha M= .816; *SD* = .082 Test-retest = .813; *SD* = .098 EI scale M= .838; *SD* = .052 SN scale M= .843; *SD* = .052 TF scale M= .764; *SD* =.122 JP scale M= .822; *SD* =.073 Median not reported	Descriptive statistics and box plots for alpha and test-retest presented.
Capraro, Capraro, & Henson (2001) Mathematics Anxiety Rating Scale	Published articles no mention of dissertations or non-English versions	17 articles 25% of the articles found	Total number of samples 35 28 alpha 7 test-retest	Alpha M= .915; *SD* = .083 Test-retest M= .841; *SD* =.073 Median not reported	4 regression models Descriptive statistics presented for these regression models
Caruso (2000) NEO personality scales	Published articles only	37 articles 15% of the articles found	Total number of samples 51 47 alpha 4 test-retest	NEO scales N scale M= .88; Md = .88; *SD* = .07 E scale M= .83; Md = .83; *SD* = .09 O scale M= .79; Md = .79; *SD* =.13 A scale M= .75; Md = .77; *SD* = .10 C scale M= .83; Md = .84; *SD* = .47	Fisher-z transformation applied Used correction for restriction of range formula Descriptive statistics reported Correlations Analysis of Variance (ANOVA)

Table 1 (continued)

Reliability Generalization Studies

Study and Instrument (or construct) investigated	Additional inclusion criteria	Number of articles found that reported reliability for the data "in hand"	Number of samples and type of reliability	Reliability Mean, Median and Standard Deviation.	Type of Analysis
Caruso & Edwards (2001) Junior Eysenck Personality Questionnaire	Published articles only Non English versions omitted Test-retest omitted	23 articles 7% of the articles found	Total number of samples 44 All were alpha	P scale M= .68; Md = .68; *SD* = .09 E scale M= .73; Md = .73; *SD* = .07 N scale M= .78; Md = .80; *SD* =.08 L scale M= .77; Md = .79; *SD* = .10	Fisher-z transformations applied Descriptive statistics for scales and predictor variables Regression Analysis
Caruso, Witkiewitz, Belcourt-Dittloff, & Gottlieb (2001) Eysenck Personality Questionnaire	Published articles only Non English versions omitted Test-retest omitted	44 articles 2.9% of the article found	Total number of samples 69 for three of the scales and 65 for one. All were alpha	P scale M= .66; Md = .68; *SD* = .13 E scale M= .82; Md = .82; *SD* = .05 N scale M= .83; Md = .83; *SD* =.04 L scale M= .77; Md = .78; *SD* = .05	Fisher-z transformations applied Descriptive statistics for scales and predictor variables Multiple regression analysis
De Ayala, Vonderharr-Carlson, & Kim (2005) Beck Anxiety Inventory scores	Published articles only Some non-English versions omitted	47 articles 32.4% of the articles found	43 alpha 12 test-retest	Alpha M= .91;*SD* = .03 Test-retest =.66; *SD* = .22 Medians not reported	Descriptive statistics Box plot Bivariate correlation analysis Alpha and test-retest analyzed separately
Deditius-Island & Caruso. (2002) Zuckerman's Sensation Seeking Scale, form V	Published articles only	21 articles 8.6% of the articles found	Total number of samples 113 All were alpha	TAS scale M= .75; Md = .75 ; *SD* = .07 ES scale M= .69; Md = .66; *SD* = .10 DIS scale M= .69; Md = .71; *SD* = .08 BS scale M= .62; Md= .61; *SD* = .16 Total M= .76; Md = .78; *SD* = .10	Fisher-z transformation applied. Descriptive statistics presented Correlation analysis
Hanson, Curry, & Bandalos (2002) Working Alliance Inventory	Published articles only	25 articles 38% of articles found	Total number of samples 73 67 alpha 6 interrater reliability (observer version)	Client M= .93; *SD* = .04 Client Short M= .95; *SD* = .03 Therapist M= .91; *SD* = .05 Therapist-Short M= .92; *SD* = .04 Observer M =.79; *SD* = .12 Medians not reported	Descriptive statistics Stem and leaf display, Box Plots Bivariate correlation analysis
Helms (1999) White Racial Identity Attitude Scale	Studies from a previous meta-analysis study	38 articles	28 alphas for all five scales 3 alphas for four scales	Contact M= .51 Disintegration M= .75 Reintegration M= .76 Pseudo M= .66 Autonomy M= .59 Median and standard deviation not reported	One-tail chi squared analysis UX test Pearson correlation analysis

Table 1 (continued)

Reliability Generalization Studies

Study and Instrument (or construct) investigated	**Additional inclusion criteria**	**Number of articles found that reported reliability for the data "in hand"**	**Number of samples and type of reliability**	**Reliability Mean, Median and Standard Deviation.**	**Type of Analysis**
Hellman, Fuqua & Worley (2006) Survey of Perceived Organization Support	Published articles from a previous meta-analysis and additional published studies found in a search	56 articles 90.3 % of articles found	Total number in the sample 77 All were alpha	Mean = .88; Md = .90; *SD* = .10	Descriptive statistics Box plot Bivariate correlation analysis
Henson & Hwang (2002) Kolb's Learning Style Inventory	Published articles only	34 articles 30.9 % of the articles found	Total number of samples 388 206 alpha 182 test-retest	Alpha CE Med .= 75; error = .25 RO Med =.79 ; error = .21 AE Med =.81; error = .19 AC Med = .80; error = .20 Test –retest CE Med .= 40; error = .60 RO Med =.52 ; error = .48 AE Med =.55; error = .45 AC Med = .56; error = .44 Means and standard deviations not reported	Descriptive statistics and Box Plots alpha and test-retest were displayed separately . Multiple regression
Henson, Kogan, & Vacha-Haase (2001) Teacher Efficacy Scale, Science Teaching Efficacy Belief Instrument, Teacher Locus of Control, and Responsibility for Student Achievement	Published articles only	52 articles 5.3% of the article found	Total number in the sample 86 All alpha	RSA+ M= .76; *SD* = .03 RSA- M= .84; *SD* = .04 TLC- I+ M= .74; *SD* = .02 TLC- I- M= .70; *SD* = .13 PSTE M= .88 ; *SD* = .05 STOE M= .761; *SD* =.025 PTE M = .778; *SD* = .057 GTE M= .696; *SD* = .072 Medians not reported	Descriptive statistics Box plots Bivariate correlations
Kieffer & Reese (2003) Geriatric Depression Scale	Published articles only	98 articles 28.99% of the article found 117 articles reported means and standard deviations that were used to calculate KR 21	Total number in sample 267 100 alpha 33 test - retest 134 calculated KR21	Overall without KR-21 estimates M= .85; *SD* = .09 Over all with KR-21 estimates M= .8027; *SD* =.14 Alpha M= .8522; *SD* = .09 Test-retest M= .83; *SD* = .08 KR-21 estimates M= .76; *SD* = .14 Medians not reported	Descriptive statistics were presented for each of the reliability types separate and together. Box Plots for the 133 coefficients Compared to the 267 (added KR21) Multiple regression separately for the 133 reliabilities and the 267 coefficients

Table 1 (continued)

Reliability Generalization Studies

Study and Instrument (or construct) investigated	**Additional inclusion criteria**	**Number of articles found that reported reliability for the data "in hand"**	**Number of samples and type of reliability**	**Reliability Mean, Median and Standard Deviation.**	**Type of Analysis**
Lane, White, & Henson (2002) Coopersmith Self-Esteem Inventory	Published articles in English only	33 articles 11.97 % of the articles found 107 articles reported means and standard deviations that were used to calculate KR 21	Total number in the sample 683 66 KR20/ alpha 69 test-retest 548 calculated KR21	KR20/alpha M= .729; *SD* =.14 Test-retest M= .55; *SD* = .172 KR-21 M= .67; *SD* =.31 Medians not reported	Descriptive statistics Scatter graph depicting the relationship between KR-21 and KR20/alpha Box plots Regression analysis ANOVA
Leach, Henson, Odom, & Cagle (2006) Self-Description Questionnaire.	Published articles only Test-retest omitted	56 articles 50% of the articles found	813 alphas Three subscales evaluated separately SDQ I (n =29), SDQ II (n =13), SDQ III (n = 24)	SDQ I M= .92; *SD* = .04 SDQ II math M= .93; *SD* = .01 Verbal M= .85; *SD* = .04 GS M= .85; *SD* = .01 GSC M= .86; *SD* = .03 SDQ III Not reported Medians not reported	Descriptive statistics Regression analysis ANOVA
Nilsson, Schmidt, & Meek (2002) Career Decision-Making Self-Efficacy Scale	Published articles and dissertations	20 articles/dissertations 41% of the articles found	Total number in the sample 20 19 alpha 1 test-retest	CDMSE Mean = .95; *SD* = .04 CMESE- short form Mean = .94; *SD* = .01 Medians not reported	Descriptive statistics Bivariate correlations ANOVA
O'Rourke, (2004) Center for Epidemiologics Studies-Depression (CES-D) Scale	Published articles and dissertation	106 articles/dissertations 68% of the articles found	Total number in the sample 141 11 test-retest 130 alpha	Mean = .88, Md = .89; *SD* = .05	Descriptive statistics presented for alpha and test-retest separately Test-retest sample (n = 11) was only evaluated using a Correlation coefficient. Descriptive statistics, box plot and regression analysis for alpha only (n = 130)
Reese, Kieffer, & Briggs (2002) Adult Attachment Scale Bell Object Relations Inventory Inventory of Parent and Peer Attachment Parental Attachment Questionnaire Parental Bonding Instrument	Published articles only	53 articles 34.4% of the articles found	Total number in the sample 53 44 alpha 9 test-retest	Combined AAS M= .75; *SD* = .07 BORI M= .77; *SD* = .08 IPPA M= . 87; *SD* = .08 .PAQ M= .89; *SD* = .05 PBI M= .82; *SD* = .11 Medians not reported	Descriptive statistics presented for alpha and test-retest separately and combined Box Plots of combined reliabilities for each subscale Bivariate correlations

Table 1 (continued)

Reliability Generalization Studies

Study and Instrument (or construct) investigated	Additional inclusion criteria	Number of articles found that reported reliability for the data "in hand"	Number of samples and type of reliability	Reliability Mean, Median and Standard Deviation.	Type of Analysis
Ross, Blackburn, & Forbes (2005) Patterns of Adaptive Learning Survey	Published articles only	30 articles 47 % if the articles found	Total number in the sample 103 alphas	Overall M= .77; *SD* = .07 EX scale M= .68; *SD* = .07 TG scale M= .79; *SD* = .05 PAP scale M= .79; *SD* = .07 PAV scale M= .81; *SD* = .04 Medians not reported	Descriptive statistics were displayed using Box Plots separating the four scales. Scales were averaged separately and together. Regression analysis
Ryngala,Shields, & Caruso (2005). Reliability Generalization of the Revised Children's Manifest Scale	Partitioned normative sample from previous study	NA	48 alphas for each of the 4 subscales from 48 sub samples	Overall M= .79; Md = .81; *SD* = .06 Phy scale M= .59; Md = .61; *SD* = .13 W&S scale M= .76; Md = .77; *SD* = .06 Consent scale M=.62; Md = 63; *SD* = .11 Lie scale M= .70; Md = .72; *SD* = .10	Fisher z transformations applied Descriptive statistics Hierarchical multiple regression analysis
Shields & Caruso (2003) Alcohol Use Disorders Identification Test	Published articles in English only	17 articles 16.3% of the articles found	Total number in the sample 24 All alpha	Mean = .79; Md = .81; *SD* = .10	Fisher z transformations applied Descriptive statistics Multiple regression analysis Hierarchical regression analysis
Shields, & Caruso (2004) Cage Questionnaire	Published English only articles	13 articles 15 of the articles found	22 alphas	Mean = .73; Md = .74; *SD* = .09	Descriptive statistics Bivariate correlation and point – biseriral correlation.
Thompson & Cook (2002) LibQUAL+^{TM} scores	The survey was administered to 20,416 persons from 43 universities in the US and Canada	NA	43 alphas from all 43 universities All alpha	Overall M= .94; *SD* = .02 S_Affect scale M= .94; *SD* = .01 Li_Place scale M= .90; *SD* = .03 Pers_Com scale M= .86; *SD* = .04 Info_Acc scale M= .72; *SD* = .07 Medians not reported	Alpha for each of the 43 university is displayed Descriptive statistics Regression analysis
Vacha-Haase(1998) Bem Sex Role Inventory	Published articles only	57 articles 9 % of the articles found	Total number in sample 87 pairs for male and female. The article reports that alpha, KR20 and test-retest were found but no "n" was reported.	Box Plots were displayed but no specific values for mean, median or standard deviations were reported	Descriptive statistics were displayed using Box Plots separating Male and female reliabilities. Alpha and test –retest are not analyzed separately. Regression analysis Canonical correlation

Table 1 (continued)

Reliability Generalization Studies

Study and Instrument (or construct) investigated	**Additional inclusion criteria**	**Number of articles found that reported reliability for the data “in hand”**	**Number of samples and type of reliability**	**Reliability Mean, Median and Standard Deviation.**	**Type of Analysis**
Vacha-Haase, Kogan, Tani, & Woodall (2001) Minnesota Multiphasic Personality Inventory (clinical)	Published articles only	153 articles 7.8 % of the articles found	10 scales had an average 49 reliability coefficients. The article reports that alpha and test-retest were found but no “n” was reported.	Hs scale M= .72; Md = .76 ;*SD* = .13 D scale M= .70; Md = .73; *SD* = .17 Hy scale M= .65; Md = .70 ; *SD* = .16 Pd scale M= .66; Md = .68 ; *SD* = .16 Mf scale M= .67; Md = .72 ; *SD* = .20 Pa scale M= .64; Md = .68 ; *SD* = .15 Pt scale M= .72; Md = .78 ; *SD* = .18 Sc scale M= .73; Md = .79 ; *SD* = .18 Ma scale M= .69; Md = .72 ; *SD* = .14 Si scale M= .81; Md = .85 ; *SD* = . 14	Descriptive statistics were displayed using Box Plots separating the 10 scales. Alpha and test –retest are not analyzed separately. Multiple regression analysis
Vacha-Haase, Tani, Kogan, Woodall, & Thompson (2001) Minnesota Multiphasic Personality Inventory (validity)	Published articles only For three of the scales L, F and K	153 articles 7.8 % of the articles found 37 articles specifically for the L, F, and K scales	47 coefficients for the L scale 48 coefficients each for the F and K scales The article reports that alpha, and test-retest were found but no “n” was is reported.	L scale M= .68; Md = .71; *SD* = .16 F scale M= .68; Md = .72; *SD* = .18 K scale M= .73; Md = .76; *SD* = .13	Descriptive statistics were displayed using Box Plots separating the three scales. Alpha and test–retest are not analyzed separately. Regression analysis
Viswesvaran & Ones (2000) “Big Five Factors”	Published technical manuals	28 technical manuals	Total number in the sample 2207 1359 alpha 848 test-retest	Alpha Emotional Stability M= .78; *SD* = .11 Extraversion M= .78; *SD* = .09 Open to Experience M= .73; *SD* = .12 Agreeableness M= .75 ; *SD* = .11 Conscientious M= .78; *SD* = .10 Test-retest Emotional Stability M= .75 ; *SD* = .10 Extraversion M= .76; *SD* = .12 Open to Experience M= .71; *SD* = .13 Agreeableness M= .69; *SD* = .14 Conscientious M= .72; *SD* = .13	Descriptive statistics and Box plots for alpha and test-retest reported separately
Wallace & Wheeler (2002) Life Satisfaction Index	Published articles only	30 articles 19.11% of the articles found	Total number in the sample 34 All alpha	Mean = .79; Md = .79; *SD* = .10	Fisher z transformations applied Descriptive statistics Bivariate Correlations

Table 1 (continued)

Reliability Generalization Studies

Study and Instrument (or construct) investigated	Additional inclusion criteria	Number of articles found that reported reliability for the data "in hand"	Number of samples and type of reliability	Reliability Mean, Median and Standard Deviation.	Type of Analysis
Yin & Fan (2000) Beck Depression Inventory	Published English only articles	90 articles 7.5% of the articles found	Total number in the study 165 142 alpha 23 test-retest 121 SEM were also calculated	Overall M= . 82; *SD* = .008 Alpha M= .84; *SD* = .007 Test-retest M= .69; *SD* = .009 Medians not reported.	Descriptive statistics for different types of reliability are reported (separately and combined) and for SEM. Eta squared calculated for an effect size Correlation analysis
Youngstrom & Green (2003) Differential Emotions Scales—IV	Secondary analysis of published studies only	None of the studies identified reported reliability. Raw data was retrieved from 30 studies	Total number in the study 30 All alpha	Fear M= .77; *SD* = .09 Self-hostility M= .74; *SD* = .15 Shyness M= .73; *SD* = .11 Sadness M= .73; *SD* = .10 Enjoyment M= .71; *SD* = .13 Anger M= .71; *SD* = .12 Guilt M= .63; *SD* = .15 Shame M= 63; *SD* = .13 Contempt M= .58; *SD* = .15 Disgust M= .61; *SD* = .11 Surprise M= .56; *SD* = .19 Interest M= .56; *SD* = .19 Negative affect M= .92; *SD* = .02 Hostility M= .77; *SD* = .08 Positive affect = .71; *SD* = .12 Medians not reported	Descriptive Statistics Biviariate correlations

Not all RG studies have been conducted on articles found through searches of published articles. For example, Helms (1999) used 38 studies from a previous meta-analysis of the White Racial Identity Scale (Helms & Carter, 1990). In their study evaluating scores for the LibQual measure, Thompson and Cook (2002) administered the scale to 20,416 persons from 43 different universities. Coefficient alpha then was calculated for all 43 universities, and these estimates were used in the RG evaluation of the scale. In their study on the Revised Children's Manifest Anxiety Scale (RCMAS; Reynolds & Paget, 1983), Ryngala, Shields, and Caruso (2005) used a normative sample derived from 13 states and 80 school districts across the United States such that they had a sample size of 4, 972 children ranging in ages between 6 and 19 years old. This information came from a study that was conducted by Reynolds and Paget (1983). Ryngala et al. (2005) included 48 subsamples (2 gender x 2 ethnic x 12 age groups = 48) from Reynolds and Paget's data for their RG study. Coefficient alpha was calculated using each of these 48 subsamples for each of the four subscales of the instrument. When Youngstrom and Green (2003) attempted to conduct a RG study on the Differential Emotions Scales IV (Izard, Libero, Putnam, & Haynes, 1993) they found no studies that reported reliability estimates. They actually contacted several authors from their search and calculated coefficient alpha from the raw data of 30 different studies. For their RG study on the Perceived Organization Support (Eisenberger, Huntington, Hutchison, & Sowa, 1986) Hellman, Fuqua, and Worley (2006) used published articles from a previous meta-analysis and additional articles found in their own search. When Viswesvaran and Ones (2002) wanted to examine the reliability of score measuring the "Big Five Factors" (Barrick & Mount, 1991) they used 28 technical manuals as a data source. Other than

Viswesvaran and Ones's study, all of the RG studies used reliability coefficients that were either calculated or reported for the actual data from the studies. Thus, the criteria for inclusion can contribute to possible bias in the statistical analysis.

Seventeen of the RG studies included both test-retest reliability and internal consistency estimates, and 15 examined only internal consistency estimates. Within most of these studies, even when both test-retest and internal consistency estimates were used, the number of studies that used internal consistency to estimate the reliability was always higher (Table 1). For example, in their study on the Marlowe-Crowne Social Desirability Scale (Crowne & Marlowe, 1960), Beretvas et al. (2002) found 93 articles that reported reliability but had a total of 182 observations, 21 of which were test-retest coefficients. As mentioned earlier several of the studies (Caruso & Edwards, 2001; Caruso et al., 2001; Leach et al., 2006) decided to omit test-retest reliabilities all together. Several of the RG studies also used a derived KR21 for dichotomously scored measures using means and construct standard deviations reported in the given studies to estimate the reliability (Henson et al., 2001; Kieffer & Reese, 2003; Lane et al., 2002).These authors argued that the use of KR21 was a possible solution to estimating reliability indices that were not given in the original study.

Many types of analysis have been used in RG studies. For almost all of the studies, descriptive statistics were available, such as the mean, sample size, and standard deviation of the scale(s). Several studies also displayed box plots (Capraro & Capraro, 2002; De Ayala et al., 2005; Hanson et al., 2002; Henson & Hwang, 2002; Hellman et al., 2006; Henson et al., 2001; Kieffer & Reese, 2003; Lane et al., 2002; Resse et al., 2002; Ross et al., 2005;Vacha-Haase, 1998; Vacha-Haase, Kogan, Tani, & Woodall, 2001;

Vacha-Haase, Tani, Kogan, Woodall, & Thompson, 2001; Viswesvaran & Ones, 2000). In their RG study on the Working Alliance Inventory (Horvath & Greenberg, 1989), Hanson et al. (2002) provided a stem-and-leaf plot of the score reliabilities. If a measure had several scales within it, the descriptive statistics were reported separately for each scale of the measure. In several studies, test-retest and coefficient alpha were analyzed together. For example, in all three of her studies (Vacha-Haase, 1998; Vacha-Haase, Kogan, et al., 2001; Vacha-Haase, Tani, et al., 2001), Vacha-Haase coded test-retest and coefficient alpha separately but did not distinguish the two when calculating descriptive statistics. In all three of these articles, box plots were used to display the distributions of reliability coefficients for each scale of the measure she was investigating. The articles did not indicate how many of the reliability coefficients were test-retest and how many were coefficient alpha. Some of the articles displayed descriptive statistics for the two types of reliabilities separately and together (Reese et al., 2002; Yin & Fan, 2000). When Kieffer and Reese (2003) and Lane et al. (2002) used data from studies to calculate KR21, they reported the descriptive statistics using all of the reliabilities together and all of the reliabilities that were not calculated from the data. In other words, the test-retest coefficients were not separated from the internal consistency reliabilities. In their study on the Coopersmith Self-Esteem Inventory (Coopersmith, 1967), Lane et al. (2002) had a total of 683 reliability coefficients, 66 were KR20/coefficient alpha, 69 were test-retest, and 548 were calculated KR21. In their article, two box plots were displayed next to each other for comparison, one without the 548 calculated KR21 and one including them. The 69 test-retest coefficients were not analyzed separately. The failure to analyze test-retest and internal consistency reliability estimates separately represents a major limitation of available RG studies. Examining reliability over time (test–retest) and examining reliability in terms of internal consistency (coefficient alpha) represent different aspects of reliability (Crocker & Algina, 1986; Henson, 2001).

For several of the studies, bivariate correlations were calculated between characteristics such as sample size and reliability, mean age and reliability, gender and reliability, and scale variance and reliability (e.g., Barnes et al., 2002; Caruso, 2000; Deditius-Island & Caruso, 2002; Hanson et al., 2002; Henson et al., 2001; Nilsson et al., 2002; Reese et al., 2002; Wallace & Wheeler, 2002; Youngstrom & Green, 2003). Some of the studies involved the use of regression analysis with reliability as the dependent variable (e.g., Capraro et al., 2001; Caruso & Edwards, 2001; Lane et al., 2002; Thomson & Cook, 2002; Vacha-Haase, 1998; Vacha-Haase, Tani, et al., 2001). When possible, many studies employed multiple regression (Caruso et al., 2001; Henson & Hwang, 2002; Kieffer & Reese, 2003; Shields & Caruso, 2003; Vacha-Haase, Kogan, et al., 2001). Only a few of the RG studies employed analysis of variance (ANOVA; Caruso, 2000; Lane et al., 2002; Nilsson et al., 2002). Only one of the studies applied mixed models (Beretvas et al., 2002).

The number of journal articles for each of the RG studies conducted so far have ranged from 14 to 153 (see Table 1). As noted earlier, there have been five RG studies that did not use articles found through searches of published articles (Helms, 1999; Ryngala et al., 2005; Thompson & Cook, 2002; Viswesvaran & Ones, 2000; Youngstrom & Green, 2003). While all of the studies indicated the number of reliabilities included and the number of articles used, only one of the RG studies actually indicated the frequency of reliabilities for each article, the RG study conducted by Beretvas et al. (2002), on the

Marlowe-Crowne Social Desirability Scale (Crowne & Marlowe, 1960). In this study they displayed a table that indicated the frequency of reliability estimates per study. These values ranged from 1 to 11. Fifty-two of the studies reported one estimate of internal consistency, 10 of the studies reported two, three of the studies reported three, four of the studies reported four, one study reported seven, one study reported eight, and one study reported 11 internal consistency estimates. Eleven of the studies reported one test-retest reliability estimate, one of the studies reported two, and one of the studies reported 11 test-retest reliability estimates.

A similar issue was seen with sample sizes. Five of the studies reported a sample size per study range (Capraro & Capraro, 2002; Caruso, 2000; Caruso & Edwards, 2001; De Ayla et al., 2005; Wallace & Wheeler, 2002). The ranges of sample sizes for the five studies were: 343 to 1078, 21 to 3856, 70 to 20,968, 40 to 1172, and 20 to 1574, respectively. Only four RG studies included information on a mean sample size (Capraro et al., 2001; Hanson et al., 2002; Lane et al., 2002; Vacha-Haase, Kogan, et al., 2001). The mean sample sizes and standard deviations from the other four studies were: *M* = 366.23, *SD* =393.04; *M* = 56, *SD* = 35; *M*= 79.33, *SD* =106.33; *M* = 81.74; *SD* =84.16, respectively. For the other RG studies, sample size information was not given.

In most studies, the magnitude of the reliabilities tended to be high, usually with means in the .80's or higher. However, studies reported reliabilities as low as -.02 (Youngstrom & Green, 2003) and as high as .98 (Wallace & Wheeler, 2002). These extreme estimates were rare, and in most cases the values ranged from approximately .40 to .90. It is important to note that only reported reliabilities were part of the sample. It is possible that the reason why the means were so high was because, in most cases, only the studies that had high reliability were published.

Issues in the Debate on Reliability Generalization

Regardless of the outcome of these studies, almost all of the studies had some discussion of the importance of reporting reliability and the problems with studies using what Vacha-Haase (1998) refers to as "reliability induction." (p. 7). Reliability induction refers to the reporting of reliability estimates from a previous study or a test manual, not from actual study data. This type of reporting is only marginally acceptable if two conditions are met. First, researchers must explicitly compare the characteristics of their samples with the characteristics of the sample from which they obtained the reliability reported (e.g., the sample used to calculate the reliability reported in the test manual). Second, the standard deviation of the scores for their sample must be similar to those from the study from which they are inducting reliability. If both of these criteria are met such that there are similarities in the sample, it would be marginally reasonable to induce reliability (Vacha-Haase, 1998).

Thompson and Vacha-Haase (2000) emphasized the need to recall reliability is based on the scores from a test and not the test itself. Dawis (1987) argues "Because reliability is a function of sample as well as an instrument, it should be evaluated on a sample from the intended population—an obvious but sometimes overlooked point" (p. 486). It is also important to note that reliability coefficients are used to correct effect sizes estimates for attenuation (Baugh, 2002) and to make inferences about the scores on the test. The APA Task Force (Wilkinson & APA Task Force on Statistical Inference, 1999) argued that "Interpreting the size of an observed effect requires an assessment of the reliability scores" (p. 596). Several measurement textbooks concur (e.g., Crocker & Algina, 1986; Gronlund & Linn, 1990; Pedhazur & Schmelkin, 1991).

Sawilowsky (2000) argued that when Thompson and Vacha-Haase (2000) refer to the reliability of the data in hand or score reliability, they are implementing "datametrics." In his article he argued, "If reliability only relates to the set of scores that a test publisher obtained in a pilot, field test, or norming procedure then what purpose do the *Mental Measurement Yearbook* and *Test in Print* serve?" (p. 117). He agreed that reliability should be reported from the researcher's sample but the reliability from the test manual also should be reported as well.

Not only are RG studies similar to validity generalization in terms of method, they also are similar in terms of publication bias. In meta-analysis, publication bias is sometimes referred to as the "file-drawer problem" (Rosenthal, 1979, p. 260). In most cases, meta-analyses are conducted using only published studies that may be biased towards statistically significant results. The missing data problem is exacerbated in RG studies because information on reliability often is not reported or the reported reliability estimates are based on instruments' technical manuals rather than based on the sample used in the research. The tendency for published research not to include estimates of score reliability yields data sources with very large proportions of missing information. For example, in their RG study of the Beck Depression Inventory (BDI; Beck et al., 1961) scores, Yin and Fan (2000) found that out of 1,200 studies that used the BDI, 80.1% (961) did not mention reliability at all, 5.6% (67) mentioned it with no citation of the estimate's source, and 6.8% (82) cited reliability from the published test manuals or other sources, leaving only 7.5% (90) of the studies that reported reliability coefficients for the data used in the actual studies. Thus, the lack of reporting of reliability coefficients for the data in hand is a common occurrence (Thompson & Snyder, 1998; Vacha-Haase, Ness, Nilsson, & Reetz, 1999). To compensate for the small amount of reported reliability estimates, some RG researchers have used KR-21 derived from reported studies to estimate reliability and increase the sample size for the RG analysis. For example, in the RG study on the Coopersmith Self-esteem Inventory (CSEI; Coopersmith, 1967), Lane et al. (2002) found 33 studies that reported some form of reliability for the data in hand; however, 107 reported sufficient descriptive information to compute a KR-21 reliability estimate. Because the CSEI is a dichotomous instrument, these authors derived 548 KR-21 coefficients to add to the pool of reliability estimates.

Some concern has been voiced regarding the use of Fisher's *z* transformation to normalize reliability estimates when conducting an RG study (Sawilowsky, 2000). Thompson and Vacha-Haase (2000) have argued that reliability coefficients are a squared metric (i.e., the squared correlation between observed scores and "true" scores) and consequently the Fisher's *z* transformation is unnecessary. This issue has been recently explored using test-rest reliability (Romano & Kromrey, 2002) and coefficient alpha (Romano & Kromrey, 2004). Results of these studies suggested the use of Fisher's *z* transformation of the reliability estimates provided a modest increase in the accuracy of the estimation of the population mean score reliability coefficient. This has also been the case in RG studies that implemented the Fisher's z transformation (Beretvas et al., 2002; Caruso et al., 2001; Wallace & Wheeler, 2002; Shields & Caruso, 2003).

With regards to the issue of sample weighting, Hunter and Schmidt (1990) developed a method in which the weighted mean correlation is computed with the individual correlations weighted in terms of their sample sizes. More weight is given to the results of studies with larger samples because these estimates have smaller sampling errors. While this method was used in Yin and Fan’s (2000) RG study, it is not common practice in RG studies. In his RG study on the NEO personality scales, Caruso (2000) addressed this issue but argued that because the sample sizes ranged from *n*= 21 to *n* = 3,856, the large samples would have much more influence than would small samples. He also stated that because he found no statistically significant correlation between sample size and reliability, sample weighting was unnecessary. Finally, he indicated that he conducted an analysis using sample size weights, and the results were no different than those obtained from the unweighted analysis. In their investigation of this issue simulating test-retest reliability estimates (2002) and internal coefficient alpha (2004) Romano and Kromrey found the use of weighted estimates provided better confidence band coverage than the use of unweighted estimates.

Interest has developed in the similarities and differences between RG analyses based on reliability coefficients and those based on the standard error of measurement or SEM. For example, in their RG study on the BDI, Yin and Fan (2000) argue that the standard error should be reported because SEM is a function of both group variability and the reliability estimate. They argued that there is not an inverse relationship between the SEM and the reliability estimate, that is, “… a lower reliability estimate does not necessarily mean the corresponding SEM will be larger” (p. 206). While Thompson and Vacha-Haase (2000) agreed that an RG study can be accomplished using the SEM, they indicated that the SEM is “rather crude” because it estimates an individual’s observed score variation in the population (i.e., holding constant the true score). When examining the distribution of the SEM, examinees that score above the mean are more likely to have a positive error of measurement and examinees that score below the mean are more likely to have a negative error of measurement. Another consideration is the further away from the mean that an individual scores on a given measure, the larger the error of measurement (Hopkins, 1998). Dimitrov (2002) also points out that relationship between reliability and SEM is based on the assumption that the error variance is the same for all scores. Finally, Thompson and Vacha-Haase (2000) pointed out that even if one chooses to use the SEM in an RG study, it can only be useful when the same scale and form is used across studies because SEM is a function of the scale. In other words, it would not make sense to look at the SEM if one was comparing studies that used different forms of a particular scale (forms with different variances) or if one was comparing multiple measures of the same construct.

Another concern with RG studies is that many of the reliabilities are not only based on different sample sizes, they are also based on different scale lengths. For example, Caruso (2000) in his study of the NEO personality scale (Costa & McCrae, 1985) used a Spearman Brown formula to adjust alpha for the different number of items. Dimitrov (2002) cautions that the split-half approach requires that the estimates have equal variances. Researchers conducting RG studies do not have access to raw data and therefore cannot test for equal variances.

An analysis of all of the previously mentioned studies suggested the generalization of the reliability of the study being analyzed was secondary. In some ways, it seemed that the purpose of these RG studies was to encourage researchers to evaluate the reliability of the measures that they employ. Most of the studies discussed in detail the percentage of studies from the articles identified that reported reliability for the data "in-hand" (see Table 1). Most of the authors also noted that even after the publication of Wilkinson and APA Task Force on Statistical Inference (1999), numerous authors still do not report the reliability of the scores in the individual studies.

Independence

Another important issue to consider is the fact that in several of the RG studies the samples used in the study did not represent independent observations. For all statistical procedures there are basic assumptions that underlie them (Glass & Hopkins, 1996; Pedhazur, 1982; Stevens, 1999). When examining mean differences (e.g., ANOVA, t-test) the main assumptions about the populations are:

1. The observations in each group are normally distributed.
2. The population variances are homogeneous (i.e., for *n* groups, $\sigma_1^2 = \sigma_2^2 = ...\sigma_n^2$)
3. The observations are independent.

These assumptions are important because the violation of any of them can lead to an increase in the probability of making a Type I or Type II error (Stevens. 1999). When statistical techniques are used to conduct research, a sample is collected to make inferences about a population. For these inferences to be tenable, the treatment of a study should comply with these assumptions. The irony is that in most research, violating these assumptions is somewhat unavoidable. Clearly, it is not possible for every set of observations in a given study to be independent and normally distributed with equal variance. As Stevens (1999) points out, the question is not "Are the assumptions being violated" but, rather, "How radically must a given assumption be violated before it has a serious effect on Type I or Type II error rates?" (p. 75). Since most meta-analyses typically involve the use of *t* and *F* tests, these assumptions are inherent in meta-analysis research. As Hedges (1982) points out, "If the assumptions for the validity of the t-test are met, it is possible to derive the properties of estimators of the effect sizes exactly" (p. 13).

Research has been conducted to investigate these assumptions (e.g., Barcikowski, 1981; Bock, 1975; Glass, Peckman, & Sanders, 1972; Kenny & Judd, 1986; Landman & Dawes, 1982; Raudenbush & Bryk, 1987; Scariano & Davenport, 1987; Tracz et al., 1992). In their literature review of the first two assumptions (i.e., normality and homogeneity of variances), Glass et al. (1972) concluded that non-normality only slighted impacts the alpha level of a study, even in cases where the distribution is skewed; given a large enough sample, the violation of the assumption of normality is not problematic (i.e., the statistical analysis is robust with larger sample sizes). The research also indicated that violating the assumption of homogeneity of variances was only problematic when group sizes are unequal such that the larger *n* divided by the smaller *n* is greater than 1.5 (Stevens, 1999).

Even though they state that violating the assumption of independence is "….far more serious…" (p 242), Glass et al. (1972) did not investigate it in their research. Stevens also argues that in regards to the assumption of independence, "…it is by far the most important assumption" (p 77). Kenny and Judd (1986) investigated the consequences of violating independence in ANOVA. Their research demonstrated that for the *F* test, the mean squared within and the mean squared between are considerably biased when nonindependence is ignored. Both Scariano and Davenport (1987) and Barcikowski (1981) investigated the impact that dependence has on the inflation of the alpha level (i.e., Type I error) of a study. The intra-class correlation (ICC) was used to measure the extent that dependence is present among observations in a study (see Definitions in Chapter 1). Both studies indicated that even when the intra-class correlation was as low as .01, the larger the number of observations in a group, the higher the Type I error rate. For example, both studies indicated that when the intra-class correlation was .01 and the number of observations within a group was 10, the actual alpha level was .06 not the assumed value of .05. When the number of observations was 100, the actual level was inflated to approximately .17. The larger intra-class correlation turned out to be even more problematic. For example, Scariano and Davenport (1987) simulated two groups of sample size, *n* = 100. When the ICC was .30, the actual alpha level was approximately .77. In other words, given a study with these characteristics, the researcher has 77% chance of making a Type I error. This happens because when observation are correlated (i.e., dependent), then the standard error is actually smaller then if they are not correlated (i.e., independent). This is an issue that should not be ignored.

Tracz et al. (1992) investigated the effect of violation of the assumption of independence when combining correlation coefficients in a meta-analysis. In their study they investigated the effect of the violation of the assumption of independence on the distribution of *r* and the distribution of correlation after a Fisher's *z* transformation. They conducted a Monte Carlo study using the following parameters: (a) sample size within a study (*n* = 20, 50, 100), (b) the number of predictors (*p* = 1, 2, 3, 5), the population intercorrelation among predictors (rho(*i*) = 0,. 30,.70), and the population correlation between predictors and criterion (rho(*p*)= 0, .03, .07). All possible combinations of these parameters were used to produce the predictors and criterion variables and the correlations between the predictor and criterion variables were calculated. The population intercorrelation was used as an index of dependence (i.e., when rho(*i*) =0 or when *p* = 1, the assumption of independence was not violated). For all the combinations of parameters and for the *r* and *z*, means, medians, and standard deviations were calculated. The Fisher's *z* transformed values of population correlation were evaluated for all combinations of parameters using 90%, 95%, and 99% confidence intervals. Their research suggested that nonindependence was not a major source of error in regards to means, medians, standard deviations, and confidence intervals.

Landman and Dawes (1982) identified five different types of violation of assumption of independence:

1. Multiple measures of outcomes obtained from the same participant within single studies;
2. Measures taken at multiple points from the same participant;
3. Nonindependence of scores within a single outcome measure. Both the complete score on the entire measure and the scores of separate scales of the measure are treated as independent;
4. Nonindependence of studies within a single article; and
5. Nonindependent samples across articles.

Considering the RG studies, two of these violations have occurred thus far in published studies: nonindependance of scores within a single measure (Caruso, 2000; Nilsson et al., 2002) and nonindependence of studies within a single article. For example, Yin and Fan's (2000) RG study on the BDI included 164 reliability coefficients from 90 studies. Similarly, Vacha-Haase's (1998) RG study on the Bem Sex Role Inventory (BSRI) used 87 reliability coefficients from 57 studies; and Caruso's (2000) RG study on the NEO personality scale used 51 reliability estimates from 37 studies. These represent clear violations of independence of observations.

There are many approaches to handling dependence of observations in a study. One approach is to ignore it. This seems to be what has been practiced in most RG studies, with multiple observations created from one study. In a RG study, the observations have characteristics in common such as the scale used but are different in the way that the observations are grouped (e.g., gender or type of reliability index). The main difficulty with this approach is that if some studies have more outcomes than others, they can influence the combined results across studies. One way that researchers have approached this problem is to weight each outcome by the inverse of the number of outcomes in a study (Becker, 2000). Although this may help in controlling for the influence that one study may have over another, it does not address dependence.

Another approach that has been recommended is sensitivity analysis (Gleser & Olkin, 1994; Greenhouse & Iyengar, 1994). This involves first analyzing the studies independently with only one outcome per study and then repeating the analysis by adding in other outcomes from each study. The idea is that if the results to the meta-analysis are similar then the dependence can be ignored.

It has also been suggested that when a study has multiple outcomes the researcher should average across the outcomes or use the median when the outcomes are parallel measures of a single construct (Raudenbush et al., 1988; Tracz et al., 1992). Rosenthal and Rubin (1986) have suggested that if a study had a rather large sample size and small differences in the inter-correlations between the outcomes, then a common composite outcome measure based on a common level of inter-correlation should be used. Similarly, Gleser and Olkin (1994) suggested deriving a composite outcome within studies by using individual intercorrelations among the outcome variables. Tracz et al. (1992) suggested that combining the statistics from nonindependent data in a correlated meta-analysis does not have a negative effect in terms of estimating means, median, standard deviations, and confidence intervals. However, they did acknowledge that violation of independence could inflate Type I error in regards to testing of mean effect sizes.

Finally, Beretevas and Pastor (2003) argued that a mixed effects model should be used to model dependence of multiple reliability estimates within a study while estimating how reliability estimates vary across or between studies. They used a three-level model where variability at the first order represented the sampling variability among estimates using a known variance. The second-level modeled the variability among samples within the same study. The third-level modeled the variability in reliability estimates among studies.

Beretevas and Pastor's study investigated the same studies that Yin and Fan (2000) used for their RG study on the Beck Depression Inventory (BDI; Beck et al., 1961). The fixed effect models determined that there were three predictors (form, student proportion, and age) that were significant. In contrast, the mixed effects model found that only two predictors were significant (student proportion and age).

Mixed–effects model also were used in the Beretevas et al. (2002) study on the Marlowe-Crowne Social Desirability Scale (MCSDS; Crowne & Marlowe, 1960). They also used Level 2 and Level 3 models and compared their results to a fixed effects model. The standard error estimates in the fixed effects model were found to be lower. They did caution researchers that smaller sample sizes (usually the case with RG studies) can have a negative impact on the estimation of the random effects at the with-in and between-studies levels. Keeping this in mind, they argued that the mixed effect does provide a better model to investigate the variability of score reliabilities.

Summary

Through the aggregation of a large number of studies, meta-analysis is a useful technique to generalize across studies. While there are many approaches to conducting a meta-analysis, these approaches also have limitations that should be considered. Over the years, many researchers have evaluated the issues surrounding several of the meta-analysis methods (e.g., Barcikowski, 1981; Bock, 1975; Glass et al., 1972; Kenny & Judd, 1986; Landman & Dawes, 1982; Raudenbush & Bryk, 1987; Scariano & Davenport, 1987; Tracz et al., 1992). The Reliability Generalization meta-analysis method has been used in 32 studies to evaluate the distribution of reliability across studies; yet, very little research has been conducted to address the possible methodological issues involving this technique. It is important to note that for the results of these RG studies to be credible, the method used to combine the results across studies must be statistically valid.

For all statistical procedures, there are basic underlying assumptions (Glass & Hopkins, 1996; Pedhazur, 1982; Stevens, 1999). When examining mean differences there are three main assumptions about the populations: observations are normally distributed, variances are homogeneous, and observations are independent. The research conducted on these assumptions has suggested that with larger sample sizes the violation of the assumption of normality does not have much of an impact on inflating Type I error rates. Similar conclusions have been made in regard to the assumption of homogeneity of variances. This assumption is only a problem when group sizes differ such that the larger n is more than 1.5 time larger than the smaller n (Stevens, 1999). The assumption of independence, however, is the most problematic. Even when the intra-class correlation is as low as .01, Type I error rates are drastically inflated (Barcikowski, 1981; Scariano & Davenport, 1987).

In general, most RG studies have violated independence by ignoring the fact that many of the score reliabilities in the sample are from the same study. Although there are many approaches that have been used to handle dependence of observations in meta-analysis, most RG researchers have chosen to ignore the dependency in their observations. In most of the RG studies each reliability coefficient is treated as independent even though it is quite common that more than one coefficient from each study was used. None of the RG studies calculated a mean or median reliability as a means to control for violation of independence. None of the RG studies investigated reliability using sensitivity analysis or chose at random a reliability estimate to represent each study. So far the only original RG study that has applied a mixed effect model is the Beretevas et al. (2002) study on the Marlowe-Crowne Social Desirabilty Scale (Crowne & Marlowe, 1960). In addition, Beretevas and Pastor (2003) used a mixed effect model method for their study that investigated the same studies that Yin and Fan (2000) used for their RG study on the Beck Depression Inventory (BDI; Beck et al., 1961). In both studies it was argued that the mixed effect model provide a better model to investigate the variability of score reliabilities.

The impact of ignoring the possible dependence in the reliability coefficients used in RG studies should be examined along with the other approaches to dealing with dependence. Thus, this research investigated the impact of violating the assumption that the observations are independent. In addition, the methods that researchers have devised to deal with dependent data in a meta-analysis also were investigated. It was expected that investigating the impact of these approaches would provide important guidelines for future RG studies such that the treatment of these studies is not compromised.

Chapter Three:

Method

This chapter outlines the experimental method and how the data were simulated to represent a typical RG study. The methodology used in the study was intended to address the stated purpose of the study.

Purpose

The purpose of this research was to examine the potential impact of selected methodological factors on the validity of conclusions from RG studies. Although all of the controversies described in Chapter 2 are important, this study focused on the issues surrounding violating the assumption that the observations are independent and the methods that researchers have devised to deal with dependent data in a meta-analysis. Factors such as (a) the magnitude of coefficient alpha, (b) sample size (i.e., number of examinees), (c) number of studies, (d) the number of reliabilities included in each journal study and (e) the intra-class correlation between journal studies (i.e., the degree of dependence between journal studies) were also considered. These factors were used in the method to investigate whether the magnitude of these factors had an impact on the accuracy of estimating reliability when four approaches to addressing the violation of independence were used: (a) treating dependent observations as independent, (b) randomly selecting a reliability index from each study, (c) calculating a mean or a median, and (d) using a two-level Mixed Effects model. In other words, for certain method factors, does violation of independence significantly impact the accuracy of estimating the true reliability parameter?

Research Questions

In RG studies the dependent variable in the analyses is the reliability estimate (Henson & Thompson, 2001). This research focused on how certain study methods, in regards to violation of independence, affect the estimated mean reliability of scores calculated across studies. The key questions that were addressed in this study were:

1. What is the effect on point and interval estimates of mean reliability of ignoring violation of independence of observations in RG studies (i.e. treating all reliability coefficients as independent observations)?
2. What is the effect on point and interval estimates of mean reliability of using a mean or median reliability from each study as part of a sample in a RG study?
3. What is the effect of randomly selecting a reliability estimate from each study as a part of a sample in a RG study?
4. What is the effect on point and interval estimates of mean reliability of using a two level mixed-effects model for RG studies (i.e. reliabilities are nested within studies)?
5. In regard to violations of independence, what impact do factors such as the magnitude of coefficient alpha, sample size, number of journal studies, number of reliability coefficients from each study, and the magnitude of the intra-class correlation (ICC) of the studies (i.e., the magnitude of the violation of independence) have when any of the methods discussed in the four research questions above are investigated?

Sample

Samples of primary studies were generated using population parameters from a three-parameter Item Response Theory (IRT) model (Table 3) that were developed by Hanson and Beguin (1999). The data in their study came from two forms, A and Z, of a

60 item American College Testing (ACT) Mathematics Assessment. The two forms did not have any items in common. Randomly equivalent groups of examinees took the assessment such that 2696 took Form A and 2670 took Form Z. These values were used to simulate scores of examinees and in turn generate coefficient alpha for various sample sizes of examinees and various test lengths.

From these simulated examinee responses, subsets of items were selected that yielded the target values of coefficient alpha. These target values, computed from the simulated examinees were used as the population values to which the subsequent sample estimates were compared.

The coefficient alpha values were generated using the information from the three-parameter model. The following table displays the number of items that were selected to simulate the population parameters:

Table 2

Number of Items Needed to Generate Reliability Parameter

ρ_{xx}	Number of Items
.30	3
.50	6
.70	11
.90	50

Table 3

Population Item Parameters Used for Simulations

		Parameters				Parameters	
Item	**A**	**B**	**C**	**Item**	**A**	**B**	**C**
1	0.642	-2.522	0.187	51	0.957	0.192	0.194
2	0.806	-1.902	0.149	52	1.269	0.683	0.15
3	0.956	-1.351	0.108	53	1.664	1.017	0.162
4	0.972	-1.092	0.142	54	1.511	1.393	0.123
5	1.045	-0.234	0.373	55	0.561	-1.865	0.24
6	0.834	-0.317	0.135	56	0.728	-0.678	0.244
7	0.614	0.037	0.172	57	1.665	-0.036	
8	0.796	0.268	0.101	58	1.401	0.117	0.057
9	1.171	-0.571	0.192	59	1.391	0.031	0.181
10	1.514	0.317	0.312	60	1.259	0.259	0.229
11	0.842	0.295	0.211	61	0.804	-2.283	0.192
12	1.754	0.778	0.123	62	0.734	-1.475	0.233
13	0.839	1.514	0.17	63	1.523	-0.995	0.175
14	0.998	1.744	0.057	64	0.72	-1.068	0.128
15	0.727	1.951	0.194	65	0.892	-0.334	0.211
16	0.892	-1.152	0.238	66	1.217	-0.29	0.138
17	0.789	-0.526	0.115	67	0.891	0.157	0.162
18	1.604	1.104	0.475	68	0.972	0.256	0.126
19	0.722	0.961	0.151	69	1.206	-0.463	0.269
20	1.549	1.314	0.197	70	1.354	0.122	0.211
21	0.7	-2.198	0.184	71	0.935	-0.061	0.086
22	0.799	-1.621	0.141	72	1.438	0.692	0.209
23	1.022	-0.761	0.439	73	1.613	0.686	0.096
24	0.86	-1.179	0.131	74	1.199	1.097	0.032
25	1.248	-0.61	0.145	75	0.786	-1.132	0.226
26	0.896	-0.291	0.082	76	1.041	0.131	0.15
27	0.679	0.067	0.161	77	1.285	0.17	0.077
28	0.996	0.706	0.21	78	1.219	0.605	0.128
29	0.42	-2.713	0.171	79	1.473	1.668	0.187
30	0.977	0.213	0.28	80	1.334	0.53	0.075
31	1.257	0.116	0.209	81	0.965	-1.862	0.152
32	0.984	0.273	0.121	82	0.71	-1.589	0.138
33	1.174	0.84	0.091	83	0.523	-1.754	0.149
34	1.601	0.745	0.043	84	1.134	-0.604	0.181
35	1.876	1.485	0.177	85	0.709	-0.68	0.064
36	0.62	-1.208	0.191	86	0.496	-0.443	0.142
37	0.994	0.189	0.242	87	0.979	0.181	0.124
38	1.246	0.345	0.187	88	0.97	0.351	0.151
39	1.175	0.962	0.1	89	0.524	-2.265	0.22
40	1.715	1.592	0.096	90	0.944	-0.084	0.432
41	0.769	-1.944	0.161	91	0.833	0.137	0.202
42	0.934	-1.348	0.174	92	1.127	0.478	0.199
43	0.496	-1.348	0.328	93	0.893	0.496	0.1
44	0.888	-0.859	0.199	94	1.215	0.867	0.076
45	0.953	-0.19	0.212	95	1.079	-0.486	0.264
46	1.022	-0.116	0.158	96	0.932	0.45	0.259
47	1.012	0.421	0.288	97	1.141	0.344	0.071
48	1.605	1.377	0.12	98	1.068	0.893	0.153
49	1.009	-1.126	0.133	99	1.217	1.487	0.069
50	1.31	-0.067	0.141	100	1.31	1.186	0.153

Found in: Hanson and Beguin (1999, April). *Obtaining a common scale for IRT item parameters using separate versus concurrent estimation in the common item nonequivalent groups equating design.*

Once the samples of examinees scores were generated, dependence was simulated by taking random samples from each simulated test administration and generating samples from populations with these reliability parameters.

Method

The research was conducted using a Monte Carlo simulation study method in which random samples were simulated under known and controlled population conditions. In the Monte Carlo study, RG studies were simulated by generating samples in primary studies, estimating reliability of scores in these samples, and then aggregating the sample reliability estimates in the RG studies. Figure 1, below, is a model for the simulation of the study.

The Monte Carlo study included five factors in the method. These factors were (a) the coefficient alpha (with ρ_{xx} = 0.30, 0.50, 0.70, and 0.90), (b) sample size in the primary studies (with average sample sizes, *n*, of 10, 50, 100, and 500), (c) number of primary studies (NPS) in the RG study (with *k* = 15, 50, 100, and 150) , (d) number of reliability estimates from each study (with *i* = 1, 2, 3, 10, and 50) and (e) the degree of violation of independence where the strength of the dependence is related to the number of reliability indices (i.e., coefficient alpha) derived from a simulated set of examinees and the magnitude of the correlation between the journal studies (with intra-class correlation ICC = 0, .0l , .30, and .90). The values chosen for each of these factors are based in part on observed factors of actual RG studies, in part on factors of the Tracz, Elmore, and Pohlmann (1992) simulation study, and mostly on values that represent a range that is reasonable and typical in RG studies.

Simulation of the data.

The research was conducted using SAS/IML version 9.1. Conditions for the study were run under Windows XP. Normally distributed random variables were generated using the RANNOR random number generator in SAS. A different seed value for the random number generator was used in each execution of the program, and the program code was verified by hand-checking results from benchmark datasets.

The target values, computed from the simulated examinees, were used as the population values to which the subsequent sample estimates were compared. For each condition investigated, several RG analyses, ranging from 1,000 to 10,000 replications, were simulated. The number of replications that was chosen for each condition varied because of the amount of time the simulations took to run on the computer. Larger values of alpha took much longer to simulate such that when alpha was .90, 10,000 replications would take over three months to simulate. In this study, 48.44% of the conditions had 1,000 replications, less than 1% had 2,000, 6.25% had 5,000, and 44.53% had 10,000. The use of 1,000 to 10,000 replications provides adequate precision for the investigation of the bias in the reliability parameter estimates. For example, 10,000 samples provide a maximum 95% confidence interval width around an observed proportion that is ± .0098 (Robey & Barcikowski, 1992).

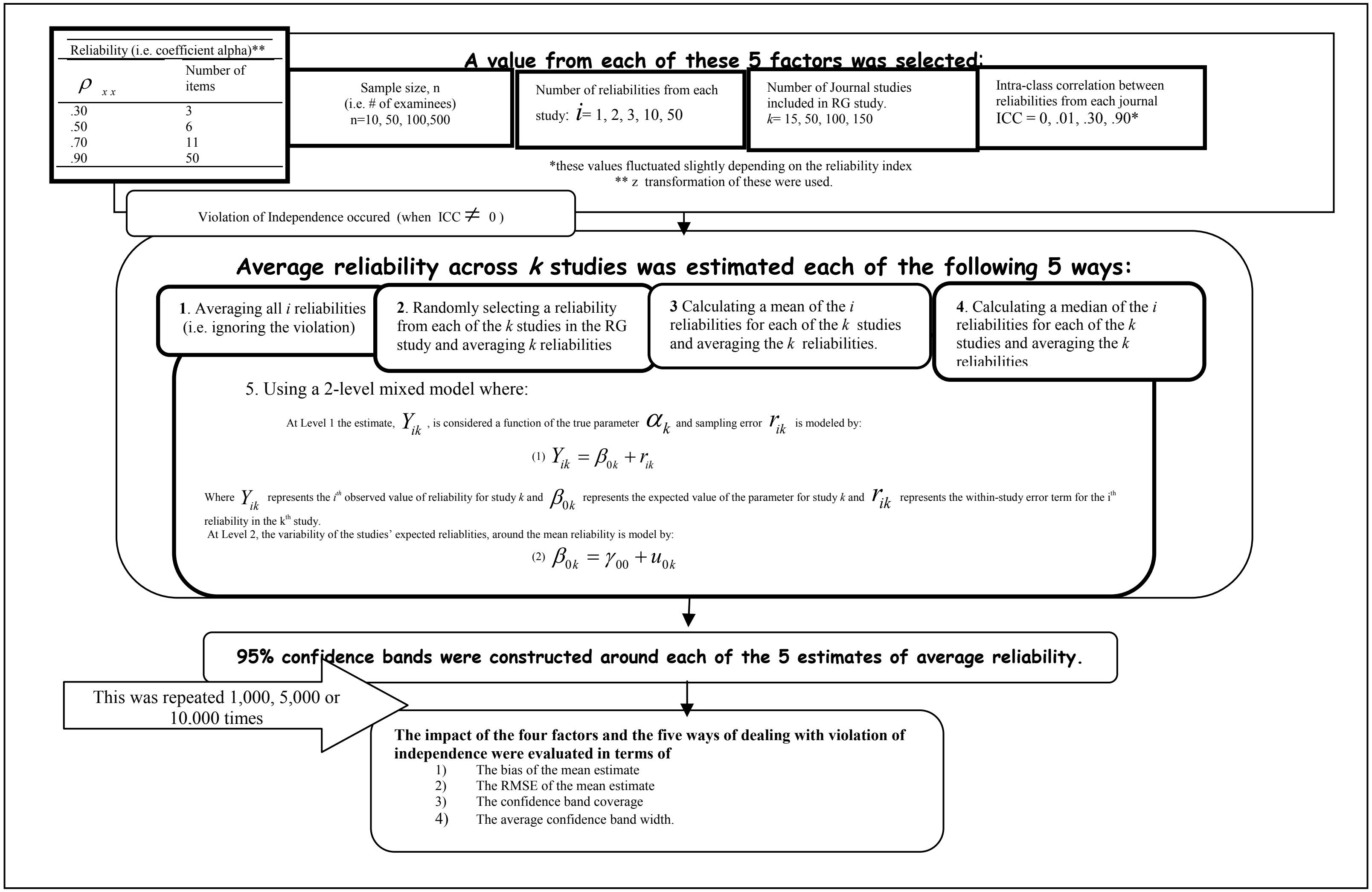

Reliability (i.e. coefficient alpha)**
ρ_{xx}
Number of items
.30 3
.50 6
.70 11
.90 50
A value from each of these 5 factors was selected:
Sample size, n (i.e. # of examinees) n=10, 50, 100,500
Number of reliabilities from each study: i= 1, 2, 3, 10, 50
Number of Journal studies included in RG study. k= 15, 50, 100, 150
Intra-class correlation between reliabilities from each journal ICC = 0, .01, .30, .90*
*these values fluctuated slightly depending on the reliability index
** z transformation of these were used.
Violation of Independence occured (when ICC $\neq$ 0)
Average reliability across k studies was estimated each of the following 5 ways:
1. Averaging all i reliabilities (i.e. ignoring the violation)
2. Randomly selecting a reliability from each of the k studies in the RG study and averaging k reliabilities
3 Calculating a mean of the i reliabilities for each of the k studies and averaging the k reliabilities.
4. Calculating a median of the i reliabilities for each of the k studies and averaging the k reliabilities
5. Using a 2-level mixed model where:
At Level 1 the estimate, Y_{ik} , is considered a function of the true parameter α_k and sampling error r_{ik} is modeled by:
(1) $Y_{ik} = \beta_{0k} + r_{ik}$
Where Y_{ik} represents the i^{th} observed value of reliability for study k and β_{0k} represents the expected value of the parameter for study k and r_{ik} represents the within-study error term for the i^{th} reliability in the k^{th} study.
At Level 2, the variability of the studies' expected reliablities, around the mean reliability is model by:
(2) $\beta_{0k} = \gamma_{00} + u_{0k}$
95% confidence bands were constructed around each of the 5 estimates of average reliability.
This was repeated 1,000, 5,000 or 10.000 times
The impact of the four factors and the five ways of dealing with violation of independence were evaluated in terms of
1) The bias of the mean estimate
2) The RMSE of the mean estimate
3) The confidence band coverage
4) The average confidence band width.

Figure 3. The model of the simulation for the study

Simulation of intra-class correlation. Intra-class correlation for coefficient alpha in the simulations was generated using the data from Hanson and Beguin (1999) Three-parameter IRT model. Recall that there were four different test lengths used to generate the four population reliability parameters (see Table 3) for this study. Basically for each simulation the number of journal studies was set to 250 and number of reliabilities within each journal study was set to 50. Each reliability coefficient that was generated for each of the journal studies was based on 2,000 administrations (i.e., 2,000 examinees) of each of the tests simulated using SAS 9.1/PROC IML. The variance within of theta each journal study was held constant at 1. The variance for alpha among journal studies was adjusted by manipulating the variance of theta (i.e., scalar ability) so that the desired levels of ICC and for the resulting set of alpha coefficients. Table 4 shows the results of these simulations.

Table 4

Results of the Simulation of Intra-class Correlation

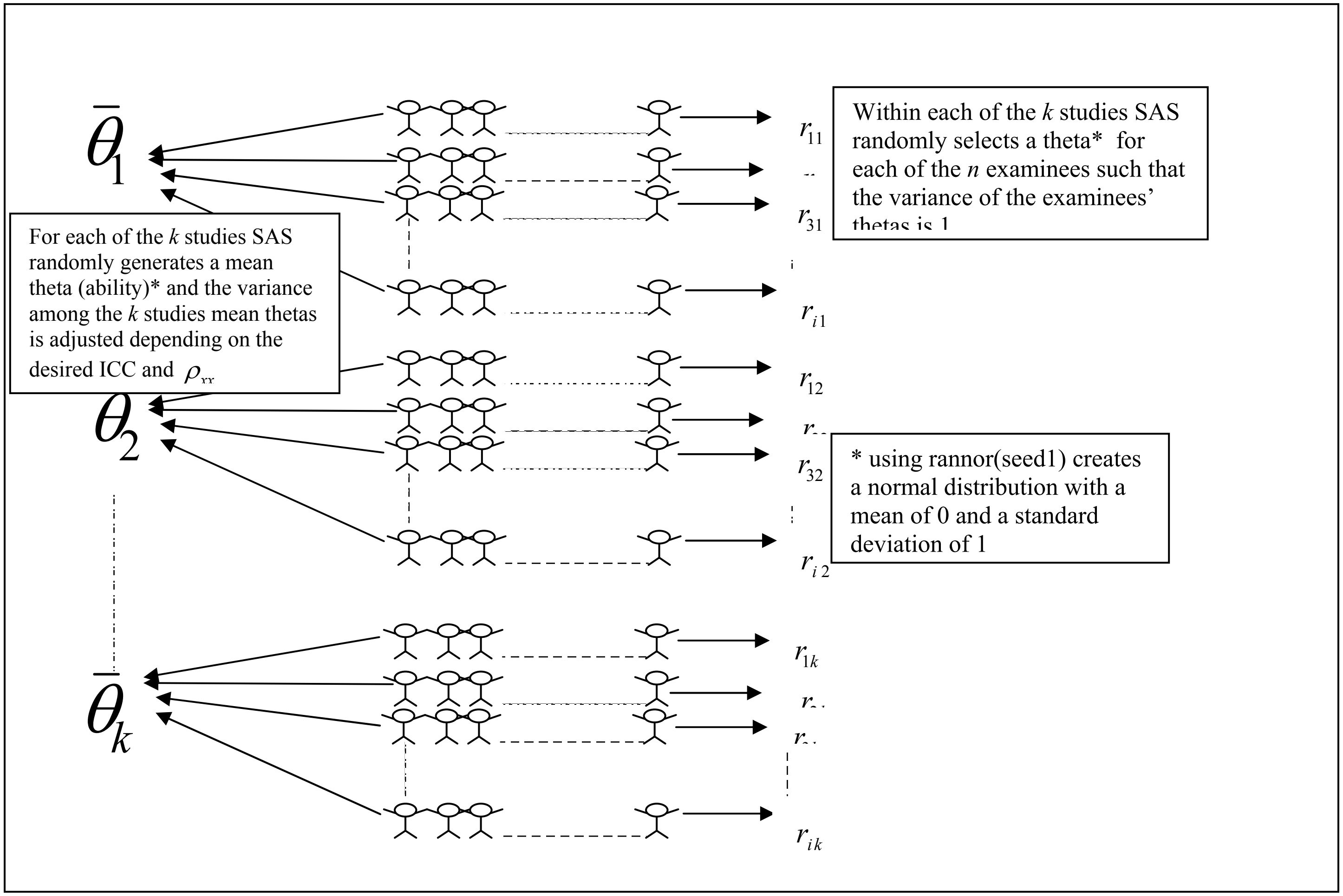
Items	Var between for theta	Var within of theta	MSb for Alpha	MSw for Alpha	ICC	Mean Alpha	What Alpha Should Be
3	0.001	1	0.00151944	0.00117117	0.01	0.33	0.30
3	0.05	1	0.02349124	0.00117762	0.27	0.33	0.30
3	3	1	0.807093	0.0020887	0.89	0.27	0.30
6	0.001	1	0.00048951	0.00030898	0.01	0.54	0.50
6	0.05	1	0.00807874	0.0003155	0.33	0.54	0.50
6	0.99	1	0.29202275	0.00059354	0.91	0.49	0.50
11	0.02	1	0.00016364	0.0000965	0.01	0.69	0.70
11	0.11	1	0.00238788	0.0000998	0..31	0.68	0.70
11	0.6	1	0.07001637	0.00013581	0.91	0.67	0.70
50	0.0005	1	0.0000218	0.00000708	0.04	0.90	0.90
50	0.005	1	0.00018575	0.00000721	0.33	0.90	0.90
50	0.1	1	0.00314615	0.0000078	0.89	0.90	0.90

Figure 2 is a model for how the “test-taking” was simulated. Once a mean theta was simulated and an examinee’s theta value was simulated then the test was administered to the examinee such that the examinee’s score was a function of the three parameters for each item and the examinee’s simulated ability level. This was repeated for each of the *n* examinees for each test administration and *i* reliabilities were generated for each of the *j* studies. For each of the RG simulations there are *j* journals and for each of the *j* journals a mean theta (i.e., ability level) was simulated from standard normal distribution. The variance between each of the *j* journal studies was fixed at a value depending on the desired intra-class correlation and coefficient alpha. For example, if a simulation was run such that the intra-class correlation was .01 and coefficent alpha was approximately .30, the variance among the mean thetas for each of the *j* journals was set to 0.001. Along with simulating a mean theta for each of the *j* studies, a theta value was simulated for each of the *n* examinees. The variance of the theta values within each admistration of the simulated test was fixed at 1.

The ICC was then generated by using the following formula

$$ICC = \frac{(MS_b - MS_w)}{(MS_b - (i-1)MS_w)}$$

where MS_b is the mean square between studies, MS_w is the mean square within studies, and *i* is the number of reliabilities for each study (Stevens, 1999).



Within each of the *k* studies SAS randomly selects a theta* for each of the *n* examinees such that the variance of the examinees' thetas is 1

For each of the *k* studies SAS randomly generates a mean theta (ability)* and the variance among the *k* studies mean thetas is adjusted depending on the desired ICC and ρ_{xx}

* using rannor(seed1) creates a normal distribution with a mean of 0 and a standard deviation of 1

Figure 4. A model for how the "test-taking" was simulated

Conduct of RG analyses. Each RG analysis was conducted using the obtained sample reliability estimates from *k* studies. Coefficient alpha estimates were investigated using the z transformation for coefficient alpha,

$$z = \ln\left(1 - |\alpha|\right)$$

to normalize the sampling distributions. This transformed value of coefficient alpha is approximately normally distributed with a variance of $k/\{2(k\text{-}1)(N\text{-}2)\}$, where k = the number of items on the instrument and N is the average sample size for each study (Bonett, 2002). Weighted least squares analyses were conducted (Fuller & Hester, 1999; Raudenbush, 1994; Hedges & Olkin, 1985).

RG analyses were conducted on the *k* studies using various approaches to address the violation of independence that were discussed in Chapter Two. First, the dependence was ignored and an RG analysis was conducted. Then, a mean and a median of the reliabilities from each of the *k* studies were calculated and an RG analysis was conducted on these averages and medians. Next, a reliability index was randomly selected from each of the *k* studies and these were the sample for an RG analysis. Finally a mixed-effects model was executed using a two-level mixed model where:

At Level 1 the estimate, Y_{ik} , is considered a function of the true parameter α_k and sampling error r_{ik} and is modeled by:

$$Y_{ik} = \beta_{0k} + r_{ik}$$

Where Y_{ik} represents the *i* observed value of reliability for study *k* and β_{0k} represents the estimated value of the parameter for study *k* and r_{ik} represents the within-study error term for the *i* reliability in the *k* study.

At Level 2, the variability of the expected reliabilities of the study, around the mean reliability is model by:

$$\beta_{0k} = \gamma_{00} + u_{0k}$$

Where β_{0k} is expressed as γ_{00}, the overall mean reliability in the k studies, and u_{0j} represents the between-study error term.

The SAS PROC MIXED procedure was used to estimate parameters from these multi-level models. The results of these approaches were evaluated in regards to their accuracy in the estimation of coefficient alpha. This was undertaken by using procedures such as PROC MEANS and PROC CORR in SAS with the output generated from the code written in PROC IML.

Evaluation of the results.

Multiple combinations of the five method factors along with the four ways of dealing with dependence within journal studies were used to simulate an RG study. Each simulated RG study was used to obtain an estimated mean reliability. In addition, a 95% confidence band was constructed around each population estimate. For the construction of confidence bands, the sampling error of each estimate of score dependability index was calculated:

$$\sigma^2_{\theta k} = \frac{\mathrm{k}}{2(\mathrm{k}-1)(\mathrm{N}-2)}$$

where $\sigma^2_{\theta k}$ is the estimated sampling variance of *z*-transformed r_{xx}

The standard error used for construction of the confidence band for the mean index of score dependability was obtained as:

$$SE_{\theta} = \left(\sqrt{\sum_{k=1}^{K} \left(\frac{1}{\sigma_{\theta k}^{2}} \right)} \right)^{-1}$$

where $\sigma_{\theta k}^{2}$ is the sampling error variance for an index θ (i.e., the transformed coefficient alpha) in the *k* study and the summation is across the studies included in the RG analysis.

The impact of the treatment factors was evaluated based upon the bias in the mean estimates, root mean square error, the confidence band coverage, and the average confidence band width. Bias was estimated as the difference between the average sample estimate and the known population value of the reliability coefficient. That is,

$$Bias\left(\hat{\theta}\right) = \frac{\sum_{i}^{R}\left(\hat{\theta}_{i} - \theta\right)}{R}$$

where $\hat{\theta}_{i}$ = the sample estimate from the *i* RG study,

θ = the population value, and the summation is over the *R* simulated RG studies.

Root mean square error estimates were calculated to evaluate the efficiency of the estimators. This value is calculated using the formula:

$$RMSE\left(\hat{\theta}\right) = \sqrt{\frac{\sum_{i}^{R}\left(\hat{\theta}_{i} - \theta\right)^{2}}{R}}$$

Confidence band coverage probabilities were estimated by computing the proportion of confidence bands in the *R* simulated RG studies that contained the parameter of interest. Similarly, confidence band width was computed as the average width of confidence bands from the *R* simulated RG studies.

Each analysis was used to obtain an estimated mean reliability and a 95% confidence band around this population estimate. Results of this research are presented as graphs of the bias, confidence band coverage, and confidence band width as functions of the method factors employed in the Monte Carlo study.

Chapter Four:

Results

The results of this study are presented in detail and are organized in the order of the research questions. The following research questions were addressed by these results:

1. What is the effect on point and interval estimates of mean reliability of ignoring violation of independence of observations in RG studies (i.e., treating all reliability coefficients as independent observations)?
2. What is the effect on point and interval estimates of mean reliability of using a mean or median reliability from each study as part of a sample in a RG study?
3. What is the effect on point and interval estimates of mean reliability of randomly selecting a reliability estimate from each study as a part of a sample in a RG study?
4. What is the effect on point and interval estimates of mean reliability of using a two level mixed-effects model for RG studies (i.e., reliabilities are nested within studies)?
5. In regard to violations of independence, what impact do factors such as the magnitude of coefficient alpha, sample size, number of journal studies, number of reliability coefficients from each study, and the magnitude of the intra-class correlation (ICC) of the studies (i.e., the magnitude of the violation of independence) have when any of the methods discussed in the four research questions above are investigated?

How the Results were Evaluated

There were 6,400 conditions simulated using the five factors of this Monte Carlo study generated from the coefficient alpha (with ρ_{xx} = 0.30, 0.50, 0.70, and 0.90), sample size in the primary studies (with average sample sizes, *n*, of 10, 50, 100, and 500), number of primary studies in the RG study (with $k = 15, 50, 100$, and 150) , number of reliability estimates from each study (with $i = 1, 2, 3, 10$, and 50), and the degree of violation of independence (ICC = 0, .01, .30, .90). In addition, the choice of treatment (ignoring the dependence, *Violation;* random selection of a reliability from each journal study, *Random*; calculating a mean from each journal study, *Mean*; calculating a median from each journal study, *Median;* and using a two-level mixed model, *HLM*) was also an independent variable for the study. Thus, this yielded 4 (ρ_{xx}) x 4(n) x 4(k) x 5(i) x 4 (ICC) x 5 (treatment) = 6,400 RG conditions. Intercorrelation analysis was conducted between the independent variables and all were equal to 0. This was because the design is a balanced factorial arrangement of factors. The results for the intercorrelation for the dependent variables are listed in Table 5. The correlation was largest in magnitude between *Bias* and *RMSE* such that $r = -.89$ and smallest in magnitude between *RMSE* and *Band Coverage* such that $r = -.17$. It was surprising to see that the correlation between *Band Coverage* and *Band Width* was only .29.

Table 5

Correlation Between Dependent Variables

	Bias	RMSE	Band Coverage	Band Width
Bias	--	-.89	.29	-.46
RMSE		--	-.17	.74
Band Coverage			--	.29
Band Width				--

Note: All correlations were significant at the α = .01 level

First, the *Bias*, root mean square error (*RMSE)*, confidence *Band Coverage*, and confidence *Band Width* were evaluated for each of the treatments. This was undertaken by creating box plots for all the conditions. These are displayed in Figures 3-6. Then, the results of the simulation were evaluated using PROC GLM in SAS such that the dependent variables were *Bias, RMSE, Band Coverage*, and *Band Width* and the independent variables were the five types of factors and the choice of treatment. The effect size, η^2 ,was calculated to measure the degree of the association between the independent variables main effects and the dependent variables along with the first-order interaction effects between the independent variables and the dependent variables. Eta-squared is the proportion of the total variance that, in the case of this study, can be attributed to one of the factors (or type of treatment) or an interaction between two of the factors (or an interaction between the type of research method and one of the factors). It is calculated as the ratio of the effect variance (SS_{effect}) to the total variance (SS_{total}).

$$\eta^2 = \frac{SS_{effect}}{SS_{total}}$$

Box Plots

To address the first four research questions, box plots were created for *Bias*, *RMSE*, *Band Coverage,* and *Band Width* to examine the results of each of the treatments. Figure 3 displays the results for the *Bias* in all five treatments. From this figure one can see that all five of the methods behave fairly the same way. The studies in which a median was calculated for each journal study do appear to have a few cases where the *Bias* was much larger in magnitude (minimum = -.27 and median = -.01), but the quartiles and the median values were similar across all five methods such that the *Bias* had a rather small range. The median value for the other types of treatment was 0. In general *Bias* was relatively very small and mostly negative, that is, the reliability from the simulations only slightly underestimated the population parameters.

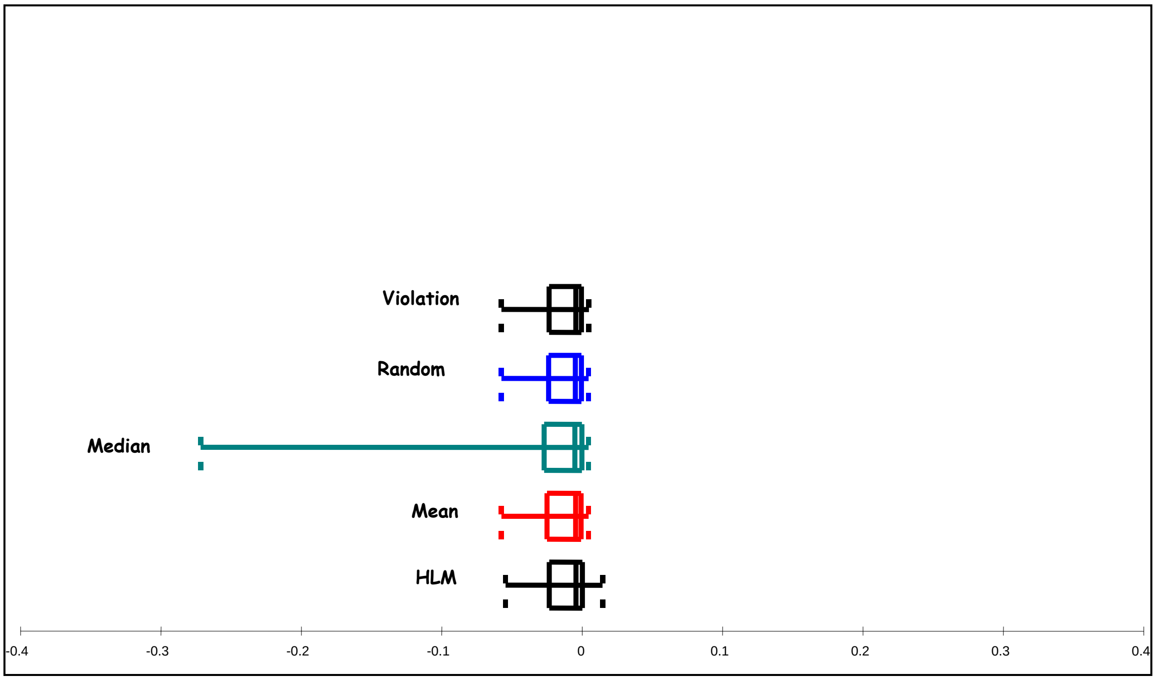


Figure 5. Distribution of bias estimates for reliability coefficients for all five types of treatments

In Figure 4 the results for *RMSE* in the study are displayed. From this figure one can see that the pattern is very similar to the results found when examining the *Bias*. The studies where a median was calculated for each journal study also appear to have a few conditions where the *RMSE* is a bit larger in magnitude (maximum = .27). The maximum value for the rest of the conditions was approximately .12. Once again, the quartiles and the medians are relatively equal for all five treatments. For all five types of treatments the minimum value was close to 0 as was the first quartile. The median value was also the same for all five treatments (median =.01). These results indicate that the estimates were very efficient regardless of the treatment.

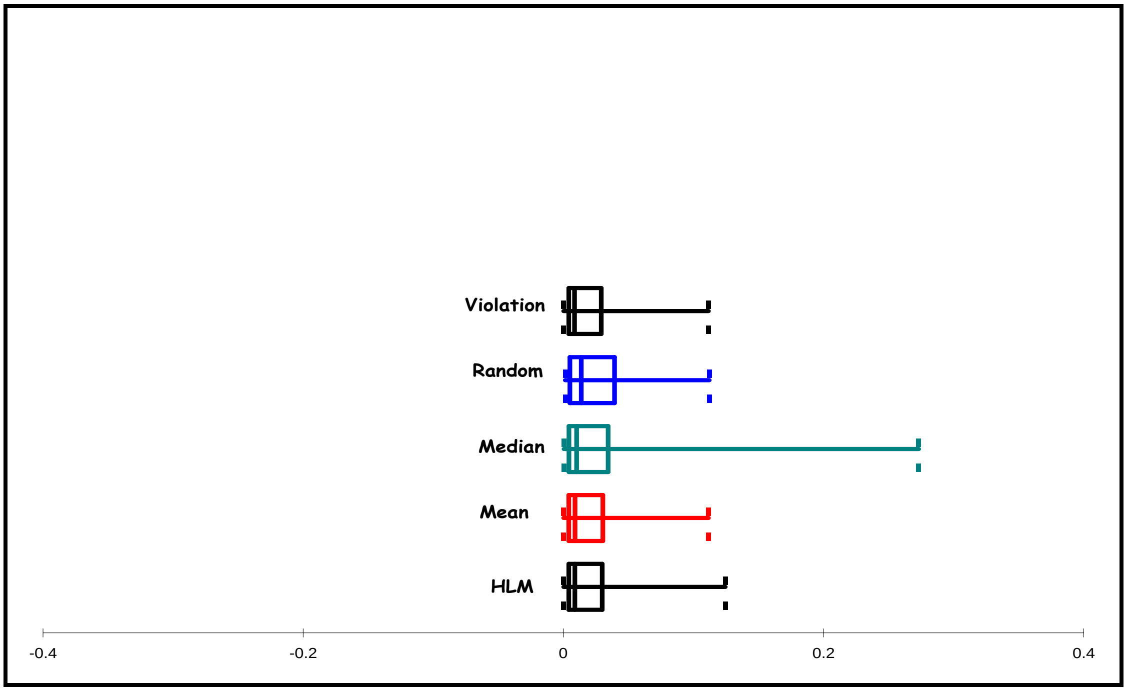


Figure 6. Distribution of RMSE estimates for reliability coefficients for all five types of treatments

In Figure 5, confidence *Band Coverage* is displayed. These values represent the proportion of times that the population reliability (i.e., coefficient alpha) fell within a 95% confidence interval for each simulation. While the range for all the treatment conditions ranged from 0 to 1, there was a wider range for simulations that ignored the dependence (*Violation*) and used mixed models (*HLM*) than the other three methods. The median *Band Coverage* for *Violation* was .54 and for *HLM* the median *Band Coverage* was .64. In contrast, the median *Band Coverage* for *Random* was .84, for *Median*, it was .89, and for *Mean* it was .94. In addition the inter-quartile range for *Violation* was .86 and for *HLM* it was .88. For the other three treatments the inter-quartile range was .44 for *Random,* .54 for *Median* and .77 for *Mean*. These results suggest calculating a mean of the reliabilities from each journal study provided better *Band Coverage* than the other four treatments. Also, these results also suggest that use of mixed models (*HLM*) provides very poor *Band Coverage* that was similar to violating independence.

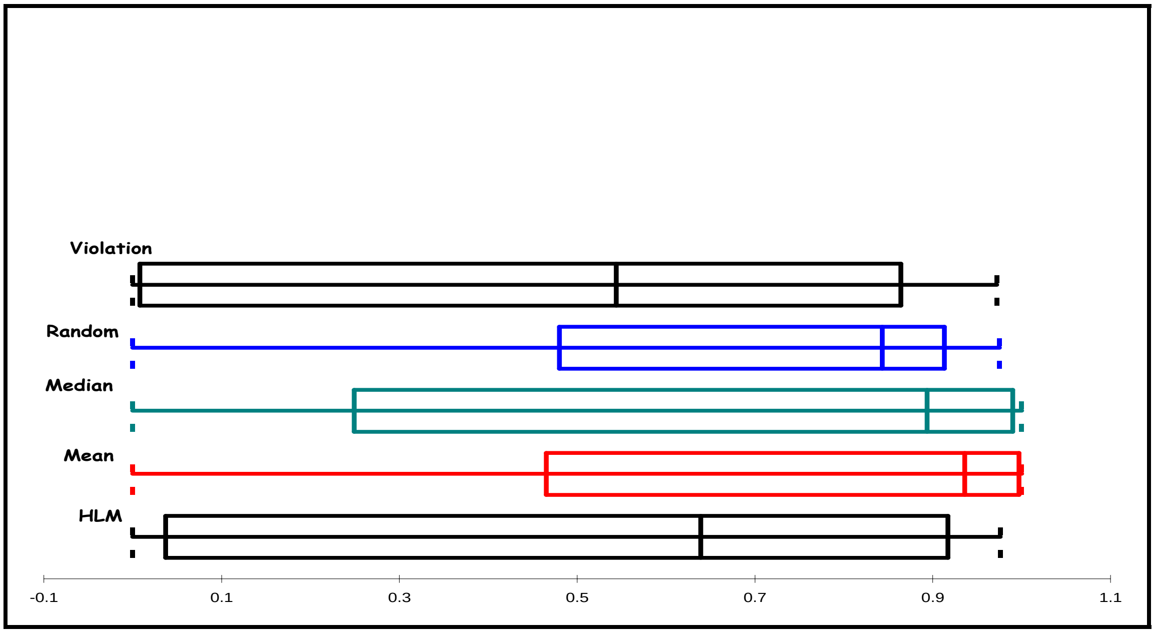
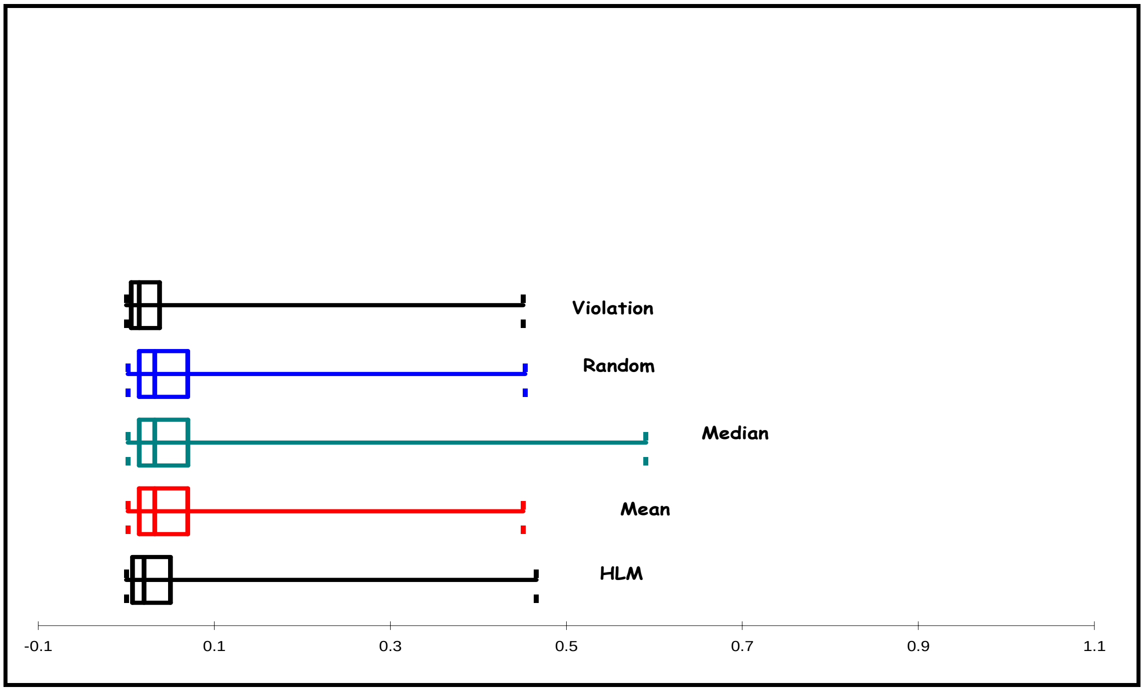


Figure 7. Distribution of band coverage for reliability coefficients for all five types of treatments

In Figure 6, the mean values of the estimated confidence *Band Width*s for reliability estimates are displayed. As with the results for *Bias* and *RMSE,* the results for all five treatments are very similar. All five methods had a minimum value of 0 and the median values for all five conditions were similar in size (.01, .03, .03, .03, and .02 for *Violation, Random, Median, Median,* and *HLM,* respectively). The largest *Band Width* was .59 and was found when median values were used in the simulation; the second largest was .47 using *HLM.* The other three types of treatment produced a maximum value of .45. As apparent in Figure 6, the inter-quartile ranges were relatively the same for all five treatments, ranging from .03 (for *Violation*) to .06 (for *Median, Mean* and

Random). While it is obvious that the inter-quartile range of the *Band Width* for *Median, Mean,* and *Random* is twice that of *Violation* and almost twice as much as that of *HLM*, (.04), these values are still small. Regardless of the treatment that was applied, the results produced very narrow bands.



F*igure 8.* Distribution of band width for reliability coefficients for all five types of treatments

Summary of Box Plot Results

The four box plot figures suggest that the five treatments used for dealing with the violation of independence do not have great impact on the variability in *Bias, RMSE* or the *Band Width*. The *Bias* across methods for the most part was relatively small and never exceeded .01. The *RMSE* analysis produced similar results. In general, the reliability estimates from the simulations only slightly underestimated the population parameter. While there is difference in the ranges of the treatments, the *Band Width* was typically rather small. There was a slightly smaller median for *Violation* and the inter-quartile range for *Mean, Median* and *Random* (.06) was twice as large as the inter-quartile range for *Violation* (.03) and almost twice as large for *HLM* (.04).

The type of treatment did seem to have an impact on the variability of *Band Coverage.* These results suggest ignoring the dependence (*Violation)* or the use of mixed models (*HLM*) provides very liberal *Band Coverage* and using a mean reliability for each study as the unit of analysis (*Mean)* seems to provide better *Band Coverage.*

η^2 *Analysis*

In addition to box plots, η^2 was calculated to measure the degree of the association between the independent variables' main effects (true alpha, average sample size from each study, number of primary studies, number of reliability estimates from each study, the degree of violation, and the treatment), and the dependent variables (*Bias*, *RMSE*, *Band Coverage*, and *Band Width*), along with the first-order interaction effects between the independent variables and the dependent variables. The results of this analysis are presented in Table 6. The η^2 values ranged from 0 to .28. According to Cohen (1988), $\eta^2 = .05$ is considered a medium effect. Using this criterion, tables and graphs were created for factors where values of η^2 were greater than or equal to .05. Even though it is clear from this analysis that the treatments for controlling for non independence only had a significant effect on *Band Coverage*, because they were addressed in the research questions, the treatments were also included in all of the analysis and presentation of the main effects results.

Table 6

η2 Analysis of the Effects of Factors in the RG Simulation

BIAS		**RMSE**		**Band Coverage**		**Band Width**	
Factor	η^2	**Factor**	η^2	**Factor**	η^2	**Factor**	η^2
ICC	0.21	ρ_{xx}	0.21	ICC	0.20	*n*	0.28
ρ_{xx}	0.17	*n*	0.19	ICC X ρ_{xx}	0.12	ρ_{xx}	0.20
N	0.12	ICC	0.14	ρ_{xx} X *n*	0.11	NPS	0.12
ICC X ρ_{xx}	0.08	ρ_{xx} X *n*	0.06	NPS	0.08	ρ_{xx} X *n*	0.10
TR X ρ_{xx}	0.05	ICC X ρ_{xx}	0.05	ρ_{xx}	0.07	n X NPS	0.07
TR X *n*	0.03	TR X ρ_{xx}	0.03	TR	0.06	ρ_{xx} X NPS	0.05
ρ_{xx} X *n*	0.02	NPS	0.02	TR X NR	0.04	TR	0.03
TR	0.02	TR X *n*	0.02	*n*	0.04	TR X *n*	0.02
TR X NR	0.01	TR	0.01	ICC X *n*	0.03	TR X NR	0.02
ρ_{xx} X NR	0.01	ρ_{xx} X NPS	0.01	NR	0.02	TR X ρ_{xx}	0.01
ICC X *n*	0.01	TR X NR	0.01	TR X ρ_{xx}	0.02	NR	0.01
NR X *n*	0.01	ICC X *n*	0.01	TR X *n*	0.01	TR X NPS	0.01
NR	0.00	*n* X NPS	0.01	ICC X NPS	0.01	NR X *n*	0.00
TR X ICC	0.00	NR X NPS	0.00	ρ_{xx} X NPS	0.01	ρ_{xx} X NR	0.00
n X NPS	0.00	NR	0.00	*n* X NPS	0.01	NR X NPS	0.00
NPS	0.00	TR X NPS	0.00	TR X ICC	0.00	ICC	0.00
ICC X NR	0.00	ρ_{xx} X NR	0.00	NR X *n*	0.00	TR X ICC	0.00
ρ_{xx} X NPS	0.00	NR X *n*	0.00	TR X NPS	0.00	ICC X ρ_{xx}	0.00
TR X NPS	0.00	ICC X NR	0.00	ρ_{xx} X NR	0.00	ICC X NPS	0.00
NR X NPS	0.00	TR X ICC	0.00	ICC X NR	0.00	ICC X *n*	0.00
ICC X NPS	0.00	ICC X NPS	0.00	NR X NPS	0.00	ICC X NR	0.00

Note. ICC = intra-class correlation, NR = number of reliability per primary journal study, NPS = number of primary studies, n = average sample size, TR = Treatment, and ρ_{xx} = coefficient alpha

Bias

Table 6 indicates that factors for *Bias* where $\eta^2 \geq 0.05$ were ICC ($\eta^2 = .21$), ρ_{xx} ($\eta^2 = .17$), *n* ($\eta^2 = .12$), the interaction between ICC and ρ_{xx} ($\eta^2 = .08$) and the interaction between the treatment and ρ_{xx} ($\eta^2 = .05$). The results using average *Bias* as an outcome and these factors as predictors are presented in Table 7 through Table 10. In addition, Figure 7 displays the interactions between ICC and ρ_{xx} in regards to *Bias* and Figure 8 displays the interaction between the treatment and ρ_{xx} in regards to *Bias*.

In Table 7 information about the extent to which the magnitude of the intra-class correlation (ICC) is associated with the *Bias* in estimated mean reliability by treatment is presented. The *Bias* was as little as approximately 0 and as large as .04 in magnitude when ICC =.90 and the treatment was *Median*. The averages of the magnitude of *Bias* for ICC ranged from .01 to .03 such that for ICC = 0, .01, and .30 the average *Bias* was -.01 and for ICC = .90 the average *Bias* was -.03. In regards to the types of treatment there was very little difference in the average *Bias*. This was of course not surprising given that the η^2 was only .02 for treatment. While the *Bias* was slightly larger for ICC =.90, it was still very small and the average *Bias* was never positive; that is, average reliability was never overestimated.

Table 7

Bias in Estimated Mean Reliability by Treatment and Intra-class Correlation

Average of BIAS	ICC				
Treatment	***0.00***	***0.01***	***0.30***	***0.90***	**Average**
Violation	-0.01	-0.01	-0.01	-0.03	-0.01
Random	-0.01	-0.01	-0.01	-0.03	-0.01
Median	-0.01	-0.01	-0.02	-0.04	-0.02
Mean	-0.01	-0.01	-0.01	-0.03	-0.01
HLM	0.00	0.00	-0.01	-0.03	-0.01
Average	-0.01	-0.01	-0.01	-0.03	-0.01

In Table 8 and in Figure 7 information about the extent to which the magnitude of the reliability parameter, ρ_{xx}, contributes to the *Bias* in estimated mean reliability by treatment is presented. The averages of the magnitude of *Bias* for ρ_{xx} ranged from 0 to .02 such that for ρ_{xx} = .33 and .54 the average *Bias* was -.02, for ρ_{xx} = .69 the average *Bias* was -.01 and for ρ_{xx} =.90 the average *Bias* was 0. While the *Bias* was slightly larger for ρ_{xx} =.33 and a *Median* treatment, it was still very small. As with the results for ICC the average *Bias* was never positive; that is, average reliability was never overestimated.

Table 8

Bias in Estimated Mean Reliability by Treatment and Coefficient Alpha

Average of BIAS	ρ_{xx}				
Treatment	**0.33**	**0.54**	**0.69**	**0.90**	**Average**
Violation	-0.02	-0.02	-0.01	0.00	-0.01
Random	-0.02	-0.02	-0.01	0.00	-0.01
Mean	-0.02	-0.02	-0.01	0.00	-0.01
Median	-0.04	-0.03	-0.01	0.00	-0.02
HLM	-0.01	-0.02	-0.01	0.00	-0.01
Average	-0.02	-0.02	-0.01	0.00	-0.01

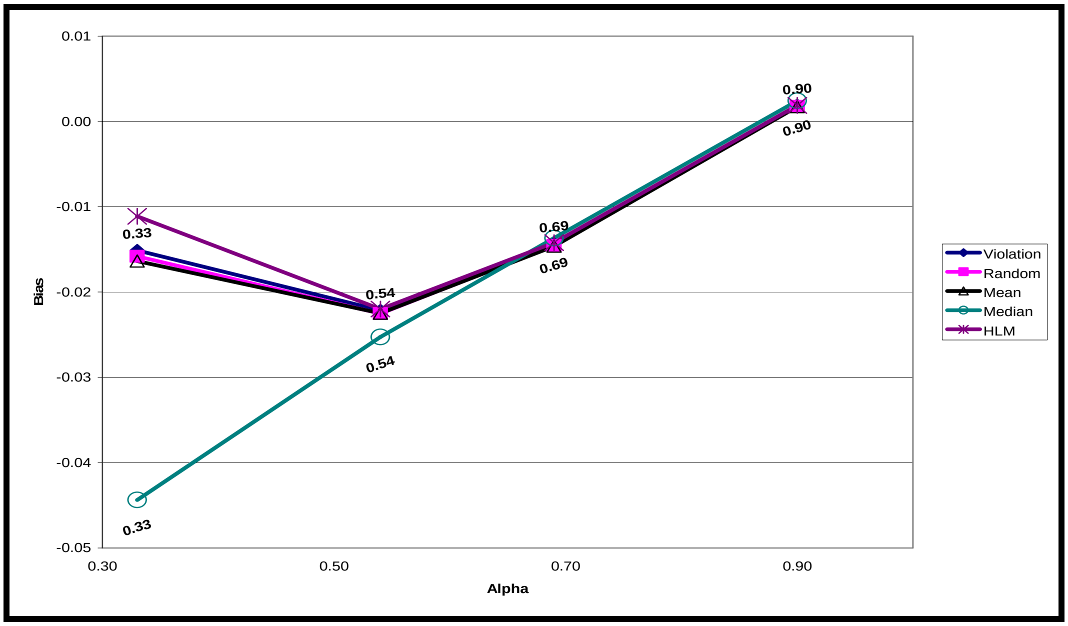


Figure 9. Bias in estimated mean reliability by treatment and coefficient alpha

In Table 9 information about the extent to which the magnitude of the average sample size, *n*, from each primary study contributes to the *Bias* in estimated mean reliability by treatment is presented. The *Bias* was as little as -.01 and as large as .05 in magnitude (-.05 when *n* = 10, and the treatment was *Median*). The averages of the magnitude of *Bias* for *n* ranged from .01 to .03 such that for *n* = 50, 100, or 500 the average *Bias* was -.01 and for *n* = 10 the average *Bias* was -.03. Like the previous results for ICC and ρ_{xx}, the magnitude of the average sample size had very little impact on the *Bias* in the estimated mean reliability.

Table 9

Bias in Estimated Mean Reliability by Treatment and Average Sample Size

Average of BIAS	**Average Sample Size**				
Treatment	**10**	**50**	**100**	**500**	**Average**
Violation	-0.02	-0.01	-0.01	-0.01	-0.01
Random	-0.02	-0.01	-0.01	-0.01	-0.01
Median	-0.05	-0.01	-0.01	-0.01	-0.02
Mean	-0.02	-0.01	-0.01	-0.01	-0.01
HLM	-0.02	-0.01	-0.01	-0.01	-0.01
Average	-0.03	-0.01	-0.01	-0.01	-0.01

Along with the interaction between treatment and coefficient alpha, another notable interaction was present between intra-class correlation and coefficient alpha. Table 10 and Figure 8 display the details of this interaction. The *Bias* in this interaction ranged from 0 to .05 in magnitude. When ICC = .90 the *Bias* was as much as five times as much as for the other smaller values of ICC considered in this study. The *Bias* for ICC = 0, .01 and .30 were relatively small and did not indicate that the magnitude of ρ_{xx} had an impact on *Bias* for these values of ICC. However, there was a notable difference when ρ_{xx}= .90. In this case, regardless of the ICC the *Bias* was zero.

Table 10

Bias in Estimated Mean Reliability by Intra-class Correlation and Coefficient Alpha

Average of BIAS			ρ_{xx}		
ICC	**0.33**	**0.54**	**0.69**	**0.90**	**Average**
0	-0.01	-0.01	-0.01	0.00	-0.01
0.01	-0.01	-0.01	-0.01	0.00	-0.01
0.30	-0.01	-0.02	-0.01	0.00	-0.01
0.90	-0.05	-0.04	-0.03	0.00	-0.03
Average	-0.02	-0.02	-0.01	0.00	-0.01

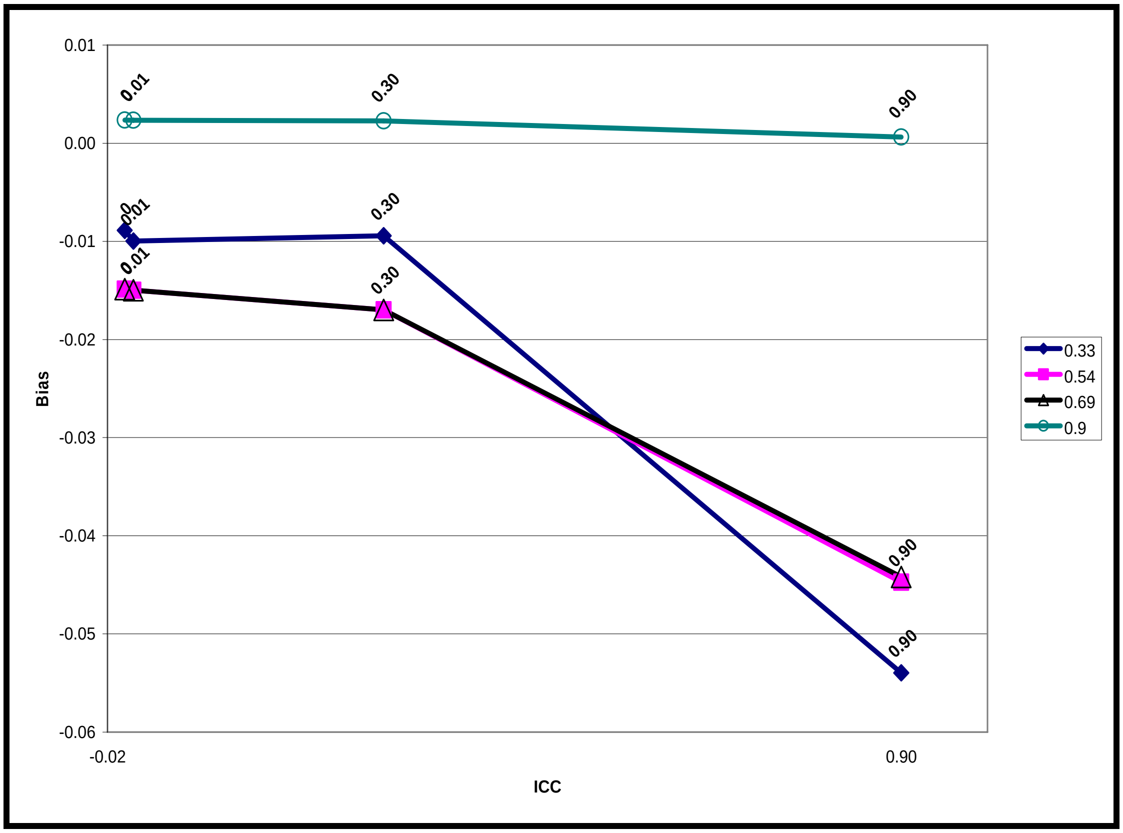


Figure 10. Bias in estimated mean reliability by intra-class correlation and coefficient alpha

Root Mean Squared Error

Table 6 indicates that factors for *RMSE* where $\eta^2 \geq 0.05$ were ρ_{xx} ($\eta^2 = .21$), *n* ($\eta^2 = .19$), ICC ($\eta^2 = .14$), the interaction between ρ_{xx} and *n* ($\eta^2 = .06$), and the interaction between ρ_{xx} and ICC ($\eta^2 = .15$). The results using average *RMSE* as an outcome and these factors as predictors are presented in Table 11 through Table 15. In addition, Figure 9 displays information about the interaction between average sample size and coefficient alpha and Figure 10 displays information about the interaction between the intra-class correlation and coefficient alpha.

In Table 11, information about the extent to which the magnitude of the reliability parameter, ρ_{xx}, contributes to the *RMSE* of estimated mean reliability by treatment is presented. The *RMSE* ranged from approximately 0 to .05. Like the results for the *Bias,* the *RMSE* was largest when $\rho_{xx} = .33$ and the treatment was *Median.* The averages of the magnitude of *RMSE* for ρ_{xx} ranged from 0 to .04 such that for $\rho_{xx} = .33$ the average *RMSE* was .04, for $\rho_{xx} = .54$ it was .03, for $\rho_{xx} = .69$ it was .02, and for ρ_{xx} =.90 the average *RMSE* was 0. These results suggest that smaller values of ρ_{xx} will have a slightly larger *RMSE* compared to larger values of ρ_{xx}. In general, the *RMSE* was quite small which would suggest that the reliability estimates were rather stable regardless of the magnitude of the population reliability parameter.

In Table 12, information about the extent to which the magnitude of the average sample size from the primary studies, *n*, contributes to the *RMSE* of estimated mean reliability by treatment is presented. The *RMSE* ranged from approximately .01 to .06. The *RMSE* was largest, .06 when *n* = 10 and the treatment was *Median.* The averages of the magnitude of *RMSE* for *n* ranged from .01 to .04 such that for *n* = 10 the average *RMSE* was .04, for *n* =50 or 100 it was .02, and for *n* = 500 the average *RMSE* was .01.

Table 11

RMSE of Estimated Mean Reliability by Treatment and Coefficient Alpha

Average of RMSE			ρ_{xx}		
Treatment	**0.33**	**0.54**	**0.69**	**0.90**	**Average**
Violation	0.03	0.03	0.02	0.00	0.02
Random	0.04	0.03	0.02	0.01	0.02
Median	0.05	0.03	0.02	0.00	0.03
Mean	0.03	0.03	0.02	0.00	0.02
HLM	0.03	0.03	0.02	0.00	0.02
Average	0.04	0.03	0.02	0.00	0.02

These results suggest that larger samples sizes have a slightly smaller and somewhat more stable *RMSE* than the smaller sample sizes. Overall, the *RMSE* was never very large which would suggest that the reliability estimates were somewhat stable regardless of the magnitude of the average sample.

Table 12

RMSE of Estimated Mean Reliability by Treatment and Average Sample Size

Average of RMSE			**Average Sample Size**		
Treatment	**10**	**50**	**100**	**500**	**Average**
Violation	0.03	0.02	0.01	0.01	0.02
Random	0.04	0.02	0.02	0.01	0.02
Median	0.06	0.02	0.02	0.01	0.03
Mean	0.03	0.02	0.01	0.01	0.02
HLM	0.03	0.02	0.01	0.01	0.02
Average	0.04	0.02	0.02	0.01	0.02

In Table 13, information about the extent to which the magnitude of the intra-class correlation contributes to the *RMSE* of estimated mean reliability by treatment is presented. The *RMSE* ranged from approximately .01 to .04. When the ICC was .90 the *RMSE* = .04 regardless of the type of treatment. For the other smaller values of ICC the variability of *RMSE* was negligible across treatments. The averages of the magnitude of

RMSE for ICC ranged from .02 to .04 such that for ICC = 0, 0.01, and .30 the average *RMSE* was .02, and for ICC = .90 the average *RMSE* was .04. These results suggest that larger values of ICC will have a larger *RMSE* than smaller values of ICC regardless of the treatment. Overall, the *RMSE* was never very large, which would suggest that the reliability estimates were somewhat stable regardless of the magnitude of the ICC.

Table 13

RMSE for Estimated Mean Reliability by Treatment and Intra-class Correlation

Average of RMSE	**ICC**				
Treatment	**0.00**	**0.01**	**0.30**	**0.90**	**Average**
Violation	0.01	0.01	0.01	0.04	0.02
Random	0.02	0.02	0.02	0.04	0.02
Median	0.02	0.02	0.02	0.04	0.03
Mean	0.01	0.01	0.01	0.04	0.02
HLM	0.01	0.01	0.01	0.04	0.02
Average	0.02	0.02	0.02	0.04	0.02

The interaction between coefficient alpha and average sample size had a significant impact on the variability in the *RMSE.* Table 14 and Figure 9 display information about these results. The *RMSE* for these data ranged from 0 to .05 in magnitude. The *RMSE* was largest (.05), when ρ_{xx} = .33 or .54 and n = 10. This was five times a much as when ρ_{xx} = .90 and n =10. As n increased *RMSE* usually decreased for any given value of ρ_{xx} , however when ρ_{xx} = .90 there was not much variability such that the *RMSE* was approximately 0 for all values of $n > 10$. In general, while there was a significant interaction between ρ_{xx} and n, the *RMSE* for the estimated mean reliability was small.

Table 14

RMSE for Estimated Mean Reliability Average Sample Size by Coefficient Alpha

Average of RMSE	ρ_{xx}				
Average Sample Size	**0.33**	**0.54**	**0.69**	**0.90**	**Average**
10	0.05	0.05	0.04	0.01	0.04
50	0.03	0.03	0.01	0.00	0.02
100	0.03	0.02	0.01	0.00	0.02
500	0.02	0.01	0.01	0.00	0.01
Average	0.03	0.03	0.02	0.00	0.02

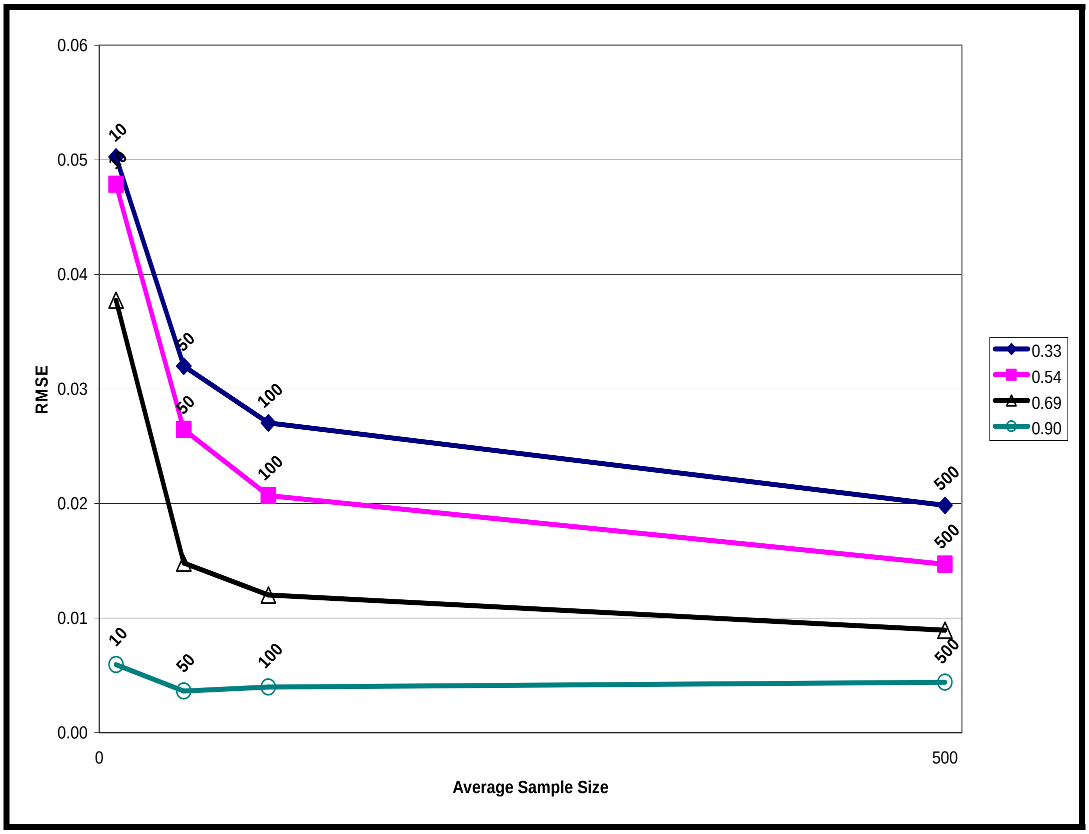


Figure 11. RMSE for estimated mean reliability average sample size by coefficient alpha

The other interaction that was significantly large for *RMSE* was the interaction between coefficient alpha and intra-class correlation. Information about this interaction is displayed in Table 15 and Figure 10. The *RMSE* was largest (.06) when ρ_{xx} = .33 and the ICC= .90. The results for this interaction were very similar to the results for the interaction between ρ_{xx} and n such that when ρ_{xx} =.90 there was very little variability in *RMSE*; actually, regardless of the value of ICC when ρ_{xx} =.90, the *RMSE* was approximately 0. For the other values of ρ_{xx} when ICC = 0, .01, or .30 the *RMSE* was relatively stable regardless of the magnitude of ρ_{xx}. When ICC = .90 and ρ_{xx} =.30 the *RMSE* was twice as large as when ICC was smaller. Overall, the larger value of ICC had the biggest impact on the magnitude of *RMSE* for smaller values of ρ_{xx} and the magnitude of ICC had no impact on the variability in *RMSE* when ρ_{xx} =.90.

Table 15

RMSE for Estimated Mean Reliability Intra-class Correlation by Coefficient Alpha

Average of RMSE			ρ_{xx}		
ICC	**0.33**	**0.54**	**0.69**	**0.90**	**Average**
0.00	0.03	0.02	0.01	0.00	0.02
0.01	0.03	0.02	0.01	0.00	0.02
0.30	0.03	0.02	0.01	0.00	0.02
0.90	0.06	0.05	0.03	0.00	0.04
Average	0.04	0.03	0.02	0.00	0.02

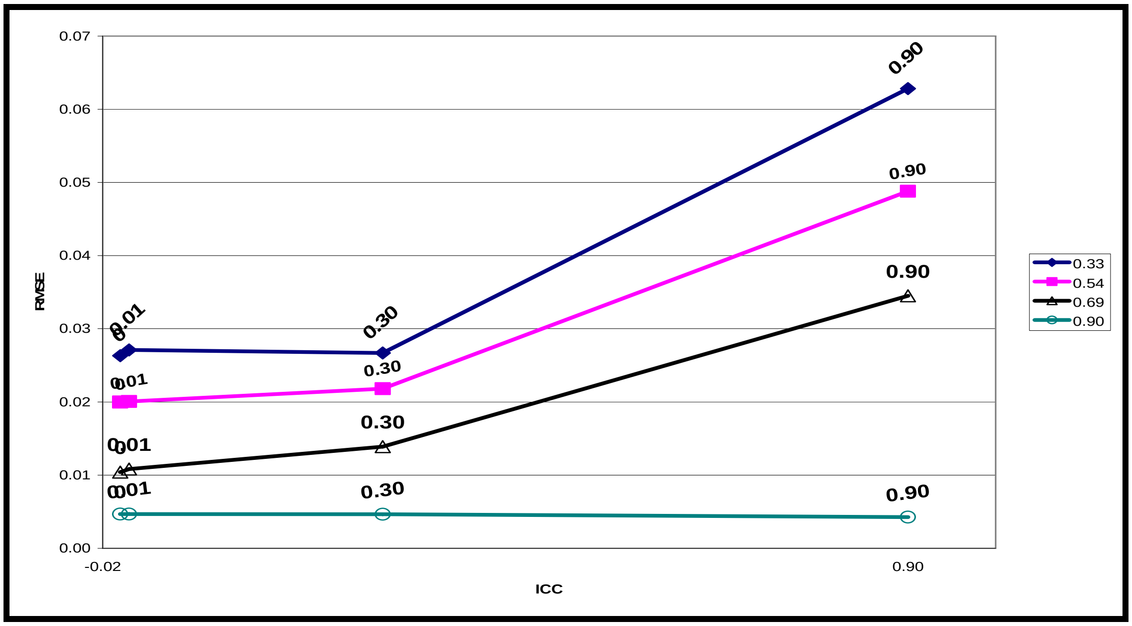


Figure 12. RMSE for estimated mean reliability intra-class correlation by coefficient alpha

Band Coverage

Table 6 indicated that factors for *Band Coverage* where $\eta^2 \geq 0.05$ were ICC ($\eta^2 = .20$), the interaction between ICC and ρ_{xx} ($\eta^2 = .12$), the interaction between ρ_{xx} and *n* ($\eta^2 = .11$), the number of primary studies ($\eta^2 = .08$), ρ_{xx} ($\eta^2 = .07$), and the treatment ($\eta^2 = .06$). Notice that, unlike *Bias, RMSE* and *Band Width*, these results indicate that the type of treatment had a notable impact on the variability in the confidence band coverage of the mean reliability estimates. These results, using average *Band Coverage* as an outcome and these factors as predictors, are presented in Table 16 through Table 20. In addition, in Figures 11 information about the interaction between ICC and ρ_{xx} is presented and in Figure 12 information about the interaction between *n* and ρ_{xx} is presented.

In Table 16 information about the extent to which the magnitude of ICC contributes to the *Band Coverage* of estimated mean reliability by treatment is presented. The *Band Coverage* ranged from approximately .20 to .85. *Band Coverage* was .20 when ICC = .90 and the treatment was *Violation* and was .85 when the ICC = 0 and the treatment was *Mean*. The averages of the magnitude of *Band Coverage* for ICC ranged from .32 to .73 such that for ICC = 0 and .01 the average *Band Coverage* was .73, for ICC = .30 it was .66, and for ICC =.90 the average *Band Coverage* was .32. These results suggest that larger values of ICC will have a much smaller *Band Coverage* compared to smaller values of ICC.

There was also some notable variability in *Band Coverage* in terms of the type of treatment. It was not surprising that out of the five treatments explored in this study, ignoring the dependence, *Violation*, had the smallest average *Band Coverage* (.47).

What was interesting was the fact that *HLM* had the second smallest average *Band Coverage* (.52) and that the largest average *Band Coverage* was for the treatment *Mean.* The type of treatment does not seem improve the size of the *Band Coverage* as ICC increases. In fact regardless of the treatment when ICC =.90 the *Band Coverage* was only as large as .38 (when the treatment was *Mean*) and as small as .20 (when the treatment was *Violation*).

Table 16

Band Coverage of Estimated Mean Reliability by Treatment and Intra-class Correlation

Average of Band Coverage	ICC				
Treatment	**0**	**0.01**	**0.30**	**0.90**	**Average**
Violation	0.60	0.60	0.49	0.20	0.47
Random	0.79	0.78	0.73	0.37	0.67
Median	0.79	0.79	0.75	0.33	0.67
Mean	0.85	0.83	0.80	0.38	0.72
HLM	0.62	0.62	0.54	0.30	0.52
Average	0.73	0.73	0.66	0.32	0.61

In Table 17 information about the extent to which the magnitude of the interaction between the intra-class correlation and coefficient alpha contributes to the variability *Band Coverage* of estimated mean reliability by treatment is presented. In Figure 11 information about the interaction between intra-class correlation and coefficient alpha also is presented. The *Band Coverage* ranged from approximately .07 to .97. *Band Coverage* was smallest, .07, when ICC = .90, ρ_{xx} =.69 and the treatment was V*iolation.* It was at its largest value, .97, twice, when ICC = 0 or and when ICC= .30, ρ_{xx} =.33 and the treatment was *Mean*. The averages of the magnitude of *Band Coverage* for ICC by ρ_{xx} ranged from .14 (when ICC = .90 and ρ_{xx} = .69), to .89, (when ICC = 0 and ρ_{xx}.= .33). Surprisingly, the range of the *Band Coverage* for ρ_{xx} = .90 was only from .42 to .61. The coverage increased for ρ_{xx} = .90 as the ICC increased with an average *Band Coverage* of .47. The average *Band Coverage* was largest, .74, for ρ_{xx} = .33.

As is apparent in Figure 11, for ρ_{xx} = .33, .54, and .69 the *Band Coverage* is fairly similar across values of ICC such that when ICC = 0 the *Band Coverage* ranges from .86 to .89 when ICC = .01, the *Band Coverage* ranges from .75 to .88, and when ICC = .30, the *Band Coverage* ranges from .67 to .87. When ICC = .90 the *Band Coverage* for these three values of ρ_{xx} drops down significantly where the *Band Coverage* ranges from .14 to .32. For ρ_{xx} =.90 a completely different pattern was seen. For this value of $\rho_{xx,}$ the *Band Coverage* was rather small, .42, and increased only when ICC = .90. Notice that this behavior was different than what was seen with the other values of ρ_{xx}. Clearly, the impact of the magnitude of coefficient alpha on *Band Coverage* depends on the magnitude of the intra-class correlation between the studies. While it was not necessarily surprising that there was an observable interaction between ICC and ρ_{xx}, it was surprising to see that *Band Coverage* for ρ_{xx} =.90 was as small as it was and that larger ICC resulted in an increase in *Band Coverage*.

Table 17

Band Coverage of Estimated Mean Reliability for Intra-class Correlation and by Coefficient Alpha

Average of Band Coverage				ρ_{xx}		
ICC	**Research Design**	**0.33**	**0.54**	**0.69**	**0.90**	**Average**
0	***Violation***	0.84	0.54	0.73	0.30	0.60
	Random	0.89	0.84	0.91	0.51	0.79
	Median	0.83	0.88	0.97	0.49	0.79
	Mean	0.97	0.93	0.96	0.52	0.85
	HLM	0.90	0.59	0.72	0.28	0.62
Average for ICC= 0		0.89	0.75	0.86	0.42	0.73
0.01	***Violation***	0.84	0.54	0.73	0.30	0.60
	Random	0.89	0.83	0.91	0.51	0.78
	Median	0.83	0.87	0.96	0.49	0.79
	Mean	0.93	0.93	0.96	0.52	0.83
	HLM	0.90	0.59	0.73	0.28	0.62
Average for ICC= .01		0.88	0.75	0.86	0.42	0.73
0.30	***Violation***	0.79	0.45	0.43	0.30	0.49
	Random	0.88	0.77	0.77	0.51	0.73
	Median	0.82	0.82	0.86	0.50	0.75
	Mean	0.97	0.87	0.84	0.53	0.80
	HLM	0.91	0.53	0.45	0.29	0.54
Average for ICC =.30		0.87	0.69	0.67	0.42	0.66
0.90	***Violation***	0.21	0.12	0.07	0.40	0.20
	Random	0.38	0.27	0.19	0.68	0.37
	Median	0.25	0.22	0.18	0.70	0.33
	Mean	0.39	0.27	0.17	0.72	0.38
	HLM	0.37	0.19	0.11	0.53	0.30
Average for ICC =.90		0.32	0.21	0.14	0.61	0.32
Average		0.74	0.60	0.63	0.47	0.61

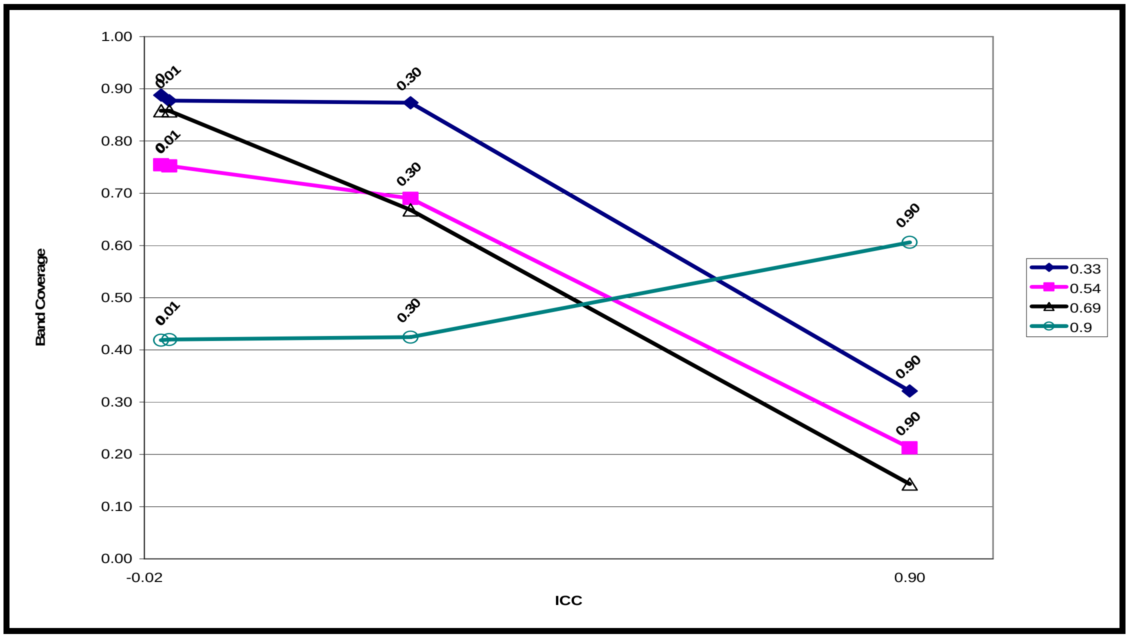


Figure 13. Band coverage of estimated mean reliability intra-class correlation by coefficient alpha

In Table 18, information about the extent to which the magnitude of the interaction between average sample size, *n*, and the population reliability parameter, $\rho_{xx,}$ contributes to the *Band Coverage* of estimated mean reliability by treatment is presented. In Figure 12 information about the extent to which the magnitude of the interaction between average sample size, *n,* and the population reliability parameter (ρ_{xx}) contributes to the *Band Coverage* of estimated of mean reliability also is presented. The *Band Coverage* ranged from approximately .03 to .98. *Band Coverage* was smallest, .03, when n = 500, ρ_{xx} =.90 and the treatment was V*iolation.* It was at its largest value, .98, twice, when ρ_{xx} =.33, n= 10 and the treatment was *Mean,* and when ρ_{xx} =.90, n = 10, and the treatment was *Median.*. The averages of the magnitude of *Band Coverage* for *n* by ρ_{xx} ranged from .06, where ρ_{xx} =.90 and n = 500, to .85, where ρ_{xx} =.90 and n = 10. The overall average of *Band Coverage* for ρ_{xx} ranged from .47, for ρ_{xx} =.90 to .74, for ρ_{xx}

=.30. The overall average *Band Coverage* for *n* ranged from .50 (for *n* = 500) to .70 (for *n* = 10).

As displayed in Figure 12, the *Band Coverage* had a wider range for smaller values of *n* such that for *n* = 10, when ρ_{xx} = .33 the average *Band Coverage* was .82, when ρ_{xx} = .54 the average *Band Coverage* was .59, when ρ_{xx} = .69 the average *Band Coverage* was .86, and when ρ_{xx}= 90 it was .62. As the average sample size increased the average *Band Coverage* for ρ_{xx} = .33, .54, and .69 did not change that drastically. This was not the case, however, for ρ_{xx} = .90. In this case, the average *Band Coverage* for *n* = 50, 100, and 500 went from .60 to .33 to .06, respectively. This would explain the interaction effect between average sample size and coefficient alpha. These results suggest that for smaller values of $\rho_{xx,}$ the *Band Coverage* is less affected by an increase in sample size than for larger values of $\rho_{xx.}$

Table 18

Band Coverage of Estimated Mean Reliability for Average Sample Size and Treatment by Coefficient Alpha

Average of Band Coverage				ρ_{xx}		
Sample Size	**Research Design**	***0.33***	***0.54***	***0.69***	***0.90***	**Average**
10	***Violation***	0.83	0.38	0.33	0.68	0.56
	Random	0.93	0.75	0.66	0.92	0.81
	Median	0.55	0.63	0.74	0.98	0.72
	Mean	0.98	0.83	0.70	0.97	0.87
	HLM	0.83	0.37	0.32	0.67	0.55
Average for Sample Size = 10		0.82	0.59	0.55	0.85	0.70
50	***Violation***	0.65	0.40	0.52	0.38	0.49
	Random	0.76	0.66	0.74	0.73	0.72
	Median	0.70	0.73	0.78	0.72	0.73
	Mean	0.77	0.73	0.78	0.80	0.77
	HLM	0.76	0.47	0.54	0.39	0.54
Average for Sample Size = 50		0.73	0.60	0.67	0.60	0.65
100	***Violation***	0.58	0.44	0.58	0.19	0.45
	Random	0.70	0.67	0.72	0.43	0.63
	Median	0.75	0.73	0.76	0.40	0.66
	Mean	0.77	0.73	0.76	0.42	0.67
	HLM	0.71	0.51	0.60	0.21	0.51
Average for Sample Size = 100		0.70	0.62	0.69	0.33	0.58
500	***Violation***	0.62	0.44	0.52	0.03	0.41
	Random	0.66	0.64	0.65	0.08	0.51
	Median	0.73	0.70	0.69	0.05	0.55
	Mean	0.74	0.70	0.69	0.05	0.55
	HLM	0.78	0.54	0.55	0.07	0.49
Average for Sample Size = 500		0.71	0.60	0.62	0.06	0.50
Average		0.74	0.60	0.63	0.47	0.61

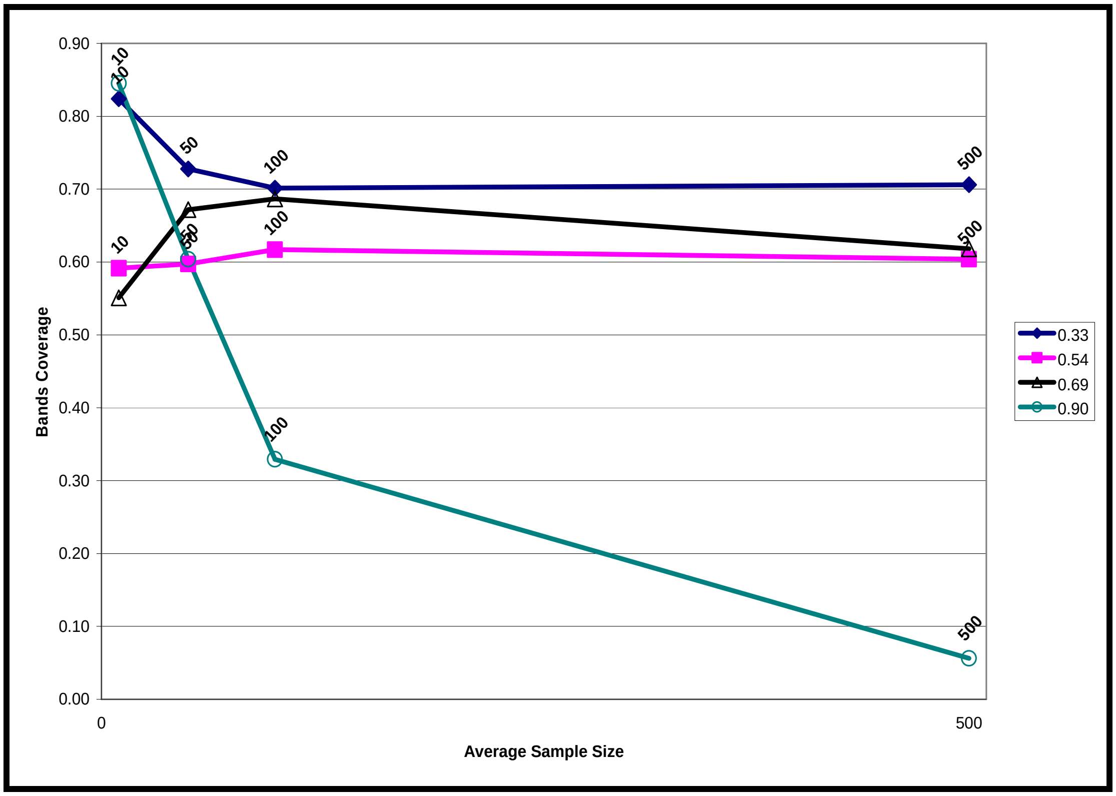


Figure 14. Band coverage of estimated mean reliability for average sample size by coefficient alpha

Unlike *Bias* and *RMSE*, the variability in *Band Coverage* was significantly influenced by the number of primary studies in the RG simulations. Table 19 displays information about *Band Coverage* in regards to the number of primary studies (NPS) and the treatments. For these conditions, the *Band Coverage* was as little as .36 (NPS= 150 and the treatment was *Violation*) and as large as .85 (NPS = 15 and the treatment was *Median*). When the NPS= 15, the *Band Coverage* was much larger than when NPS was 150. In fact when NPS = 15 and the treatment was *Violation* the *Band Coverage* was .63, however, when NPS = 150 and the treatment was *Violation*, the *Band Coverage* was almost half as much at .36. A similar pattern was seen for the other research methods such that as the NPS increased the *Band Coverage* decreased.

Table 19

Band Coverage of Estimated Mean Reliability for Number of Primary Studies by Treatment

Average of Band Coverage	**Number of Primary Studies**				
Treatment	**15**	**50**	**100**	**150**	**Average**
Violation	0.63	0.49	0.41	0.36	0.47
Random	0.80	0.70	0.61	0.56	0.67
Median	0.85	0.74	0.67	0.60	0.71
Mean	0.83	0.70	0.60	0.53	0.66
HLM	0.73	0.54	0.43	0.38	0.52
Average	0.77	0.63	0.54	0.49	0.61

Finally, the magnitude of coefficient alpha also had a significant impact on the variability in *Band Coverage.* This information is displayed in Table 20. In this situation the *Band Coverage* ranged from .32, when ρ_{xx} = .90 and the treatment was *Violation* to .81, when ρ_{xx} = .90 and the treatment was *Mean*. Overall, as ρ_{xx} increased the *Band Coverage* decreased. When the treatment was *Violation* and ρ_{xx} = .90 the *Band Coverage* (.32) was more than half the size than when ρ_{xx} = .33 and the treatment was *Violation.* A similar result was seen when the treatment was *HLM*; when ρ_{xx} = .90 the *Band Coverage* was .34 but when ρ_{xx} = .33 it was .77, more than twice as much. Obviously, the magnitude of ρ_{xx} has an impact on *Band Coverage* especially when dependence is ignored or when mixed models are applied.

Table 20

Band Coverage of Estimated Mean Reliability for Coefficient Alpha by Treatment

Average of Band Coverage	ρ_{xx}				
Treatment	**0.33**	**0.54**	**0.69**	**0.90**	**Average**
Violation	0.67	0.41	0.49	0.32	0.47
Random	0.76	0.68	0.69	0.54	0.67
Median	0.68	0.70	0.74	0.54	0.66
Mean	0.81	0.75	0.73	0.56	0.71
HLM	0.77	0.47	0.50	0.34	0.52
Average	0.74	0.60	0.63	0.46	0.61

Band Width

Table 6 indicates that factors for *Band Width* where $\eta^2 \geq 0.05$ are *n* ($\eta^2 = .28$), ρ_{xx} ($\eta^2 = .20$), NPS ($\eta^2 =$**.**12), the interaction between ρ_{xx} and *n* ($\eta^2 =.10$), the interaction between *n* and NPS ($\eta^2 =.07$), and the interaction between ρ_{xx} and NPS ($\eta^2 = .05$). Notice that this dependent variable, average *Band Width*, had the largest η^2 in this analysis. Also, note that for the other three dependent variables, *Bias, RMSE,* and *Band Coverage*, the η^2 value for ICC was always larger than .05. In contrast, for *Band Width* the η^2 for ICC was approximately 0. The results using average *Band Width* an outcome and these factors as predictors are presented in Table 21 through Table 26. In addition, the interaction between ρ_{xx} and *n* is displayed in Figure 13, the interaction between *n* and NPS is displayed in Figure 14, and the interaction between ρ_{xx} and NPS is displayed in Figure 15.

In Table 21 information about the extent to which the magnitude of the average sample size, *n*, contributes to the *Band Width* of estimated mean reliability by treatment is presented. The average *Band Width* ranged from .01 (where *n* = 500 and the treatment was *Violation*) to .14 (where *n* = 10 and the treatment was *Median*). The overall average *Band Width* for sample size ranged from .02 to .11, such that when *n* = 500, the average *Band Width* was .02, for *n* = 100, it was .03, for *n*= 50 it was .05 and for *n* = 10 it was .11. It was not surprising that there was an inverse relationship between average sample size and *Band Width* given that standard error is a function of sample size and also has an inverse relationship; that is, all things being equal, the larger the average sample size the smaller the standard error. Even though there was some variability across average sample size, overall the *Band Width* was relatively small; that is, the confidence bands were on average very narrow.

Table 21

Band Width of Estimated Mean Reliability by Treatment and Average Sample Size

Average of Band Width	Average Sample Size				
Treatment	**10**	**50**	**100**	**500**	**Average**
Violation	0.07	0.03	0.02	0.01	0.03
Random	0.13	0.05	0.04	0.02	0.06
Mean	0.13	0.05	0.04	0.02	0.06
Median	0.14	0.05	0.04	0.02	0.06
HLM	0.08	0.04	0.03	0.02	0.04
Average	0.11	0.05	0.03	0.01	0.05

In Table 22 information about the extent to which the magnitude of the population reliability parameter, ρ_{xx}, contributes to the *Band Width* of estimated mean reliability by treatment is presented. The average *Band Width* ranged from .01, where ρ_{xx} = .90 for all treatments, to .12, where ρ_{xx} = .33 and the treatment was *Median*. The overall average *Band Width* for ρ_{xx} ranged from .01 to .10, such that when ρ_{xx} =.90, the average *Band Width* was .01, for ρ_{xx} = .69, it was .04, for ρ_{xx} = .54 it was .06, and for ρ_{xx} = .33 it was .10. These results suggest that there is an inverse relationship between the magnitude of the population reliability parameter and the average *Band Width*.

Table 22

Band Width of Estimated Mean Reliability by Treatment and Coefficient Alpha

Average of Band Width	ρ_{xx}				
Treatment	**0.33**	**0.54**	**0.69**	**0.90**	**Average**
Violation	0.06	0.04	0.02	0.01	0.03
Random	0.11	0.07	0.05	0.01	0.06
Mean	0.11	0.07	0.05	0.01	0.06
Median	0.12	0.07	0.05	0.01	0.06
HLM	0.08	0.05	0.03	0.01	0.04
Average	0.10	0.06	0.04	0.01	0.05

In Table 23 information about the extent to which the magnitude of the number of primary studies contributes to the mean reliability by treatment is presented. The average *Band Width* ranged from .02, where NPS = 150 and the treatment was *Violation* to .11, where NPS = 15 and the treatments were, *Random, Median,* and *Mean*. The overall average *Band Width* for NPS ranged from .03 to .09, such that when NPS = 150 or 100, the average *Band Width* was .03, for NPS = 50, it was .05, and for NPS = 15 it was .09. As with the results for average sample size and population reliability parameter, these results suggest that there is an inverse relationship between the magnitude of the number primary studies in each RG study and the average *Band Width*.

Table 23

Band Width of Estimated Mean Reliability by Treatment and Number of Primary Studies

Average of Band Width	**Number of Primary Studies**				
Treatment	**15**	**50**	**100**	**150**	**Average**
Violation	0.06	0.03	0.02	0.02	0.03
Random	0.11	0.06	0.04	0.03	0.06
Mean	0.11	0.06	0.04	0.03	0.06
Median	0.11	0.06	0.04	0.03	0.06
HLM	0.08	0.04	0.03	0.02	0.04
Average	0.09	0.05	0.03	0.03	0.05

In Table 24 and Figure 13 information about the extent to which the interaction between average sample size, *n*, and the population reliability parameter, ρ_{xx}, contributes to the *Band Width* of estimated mean reliability is presented. The average of the magnitude of *Band Width* for *n* by ρ_{xx} ranged from 0, where ρ_{xx} =.90 and *n* = 500, to .21, where ρ_{xx} =.33 and *n* = 10. As displayed in Figure 19, while the *Band Width* for ρ_{xx} = .33 was always larger than the other values of ρ_{xx}, it had a wider range for smaller values of *n*; specifically, for *n* = 10*,* the *Band Width* ranged from .02 to .21, but for *n* = 500 the

Band Width only ranged from approximately 0 to .03. These results concur with the separate results for *n*, presented in Table 20 and for ρ_{xx}, presented in Table 21 such that these results suggest an inverse relationship between *Band Width* and these two predictors. However the additional interaction suggests that there is less variability for larger values of *n* and larger values of ρ_{xx}.

Table 24

Band Width of Estimated Mean Reliability for Average Sample Size by Coefficient Alpha

Average of Band Width	ρ_{xx}				
Average Sample Size	**0.33**	**0.54**	**0.69**	**0.90**	**Average**
10	0.21	0.13	0.08	0.02	0.11
50	0.09	0.05	0.03	0.01	0.05
100	0.06	0.04	0.02	0.01	0.03
500	0.03	0.02	0.01	0.00	0.01
Average	0.10	0.06	0.04	0.01	0.05

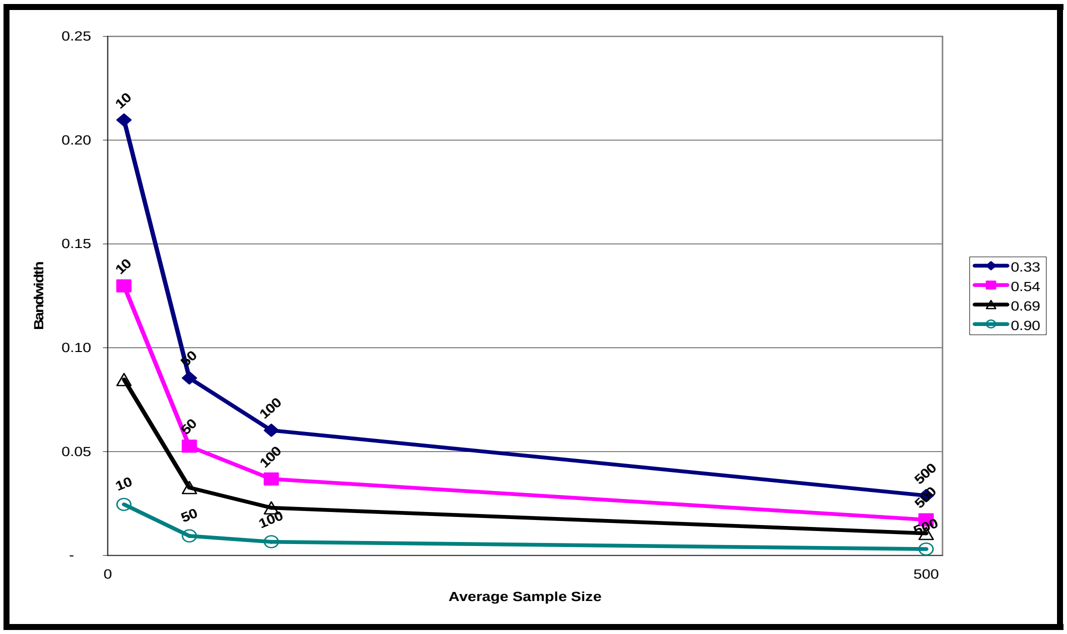


Figure 15. Band width of estimated mean reliability for average sample size by coefficient alpha

The interaction between average sample size and number of primary studies also had a significant impact on the variability in *Band Width*. This information is displayed in Table 25 and in Figure 14. For this interaction the *Band Width* ranged from .01 when *n* = 500 and NPS = 50, 100, or 150 to .20 when *n* = 10 and NPS = 15. The results for this interaction were similar to those seen for the interaction between *n* and ρ_{xx} such that smaller values of *n* had larger *Band Width*. In addition, when *n* =10 there was more variability in *Band Width* across the values of NPS than when *n*=500. For example, when *n* = 10 and NPS = 15, the *Band Width* was .20 and when NPS = 150 the *Band Width* was .06, almost one-fourth of the size. In contrast, when *n* = 500 and NPS = 15 the *Band Width* was .03 and when NPS =150 the *Band Width* was .01, one third the size.

Table 25

Band Width of Estimated Mean Reliability for Number of Primary Studies by Average Sample Size

Average of Band Width	**Average Sample Size**				
Number of Primary Studies	**10**	**50**	**100**	**500**	**Average**
15	0.20	0.08	0.06	0.03	0.09
50	0.11	0.04	0.03	0.01	0.05
100	0.08	0.03	0.02	0.01	0.03
150	0.06	0.03	0.02	0.01	0.03
Average	0.11	0.05	0.03	0.01	0.05

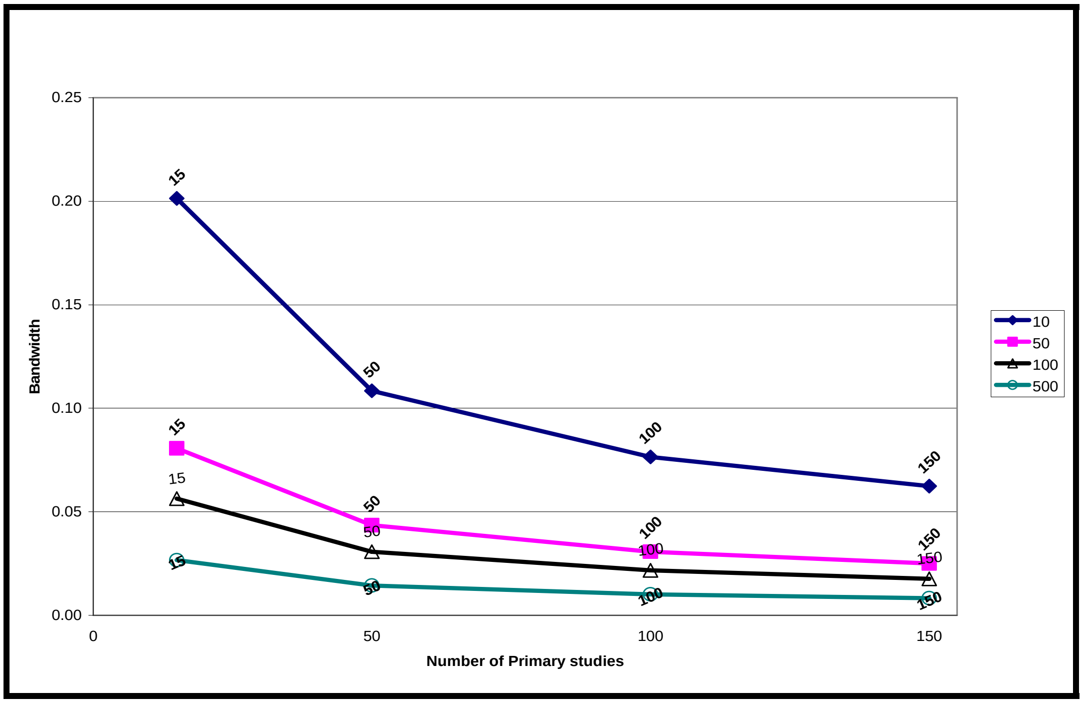


Figure 16. Band width of estimated mean reliability for number of primary studies by average sample size

The interaction between coefficient alpha and the number of primary studies also was significant in regards to the *Band Width*. The information for these results is displayed in Table 26 and in Figure 15. For this interaction the *Band Width* ranged from .01 when ρ_{xx} = .90 and NPS = 50, 100, or 150 to .17 when ρ_{xx} = .33 and NPS = 15. These results were very similar to the results for the interaction between ρ_{xx} and *n* and the interaction between *n* and NPS; smaller values produce wider confidence bands. In addition, when ρ_{xx} = .33 there was much more variability across the values of NPS than when ρ_{xx} =.90. For example, when ρ_{xx} = .33 and NPS =15 the *Band Width* was .17 and when NPS = 150 the *Band Width* was .05, almost one third the size. When ρ_{xx} = .90 and NPS =15 the

Band Width was .02 and when NPS was any other value for ρ_{xx}= .90 the *Band Width* was .01, about one-half the size. Also note that when ρ_{xx} = .33 and NPS =15, the *Band Width,* .17, is 17 times larger than the *Band Width* when ρ_{xx} = .90 and NPS =15.

Table 26

Band Width of Estimated Mean Reliability for Number of Primary Studies by Coefficient Alpha

Average of Band Width	ρ_{xx}				
Number of Primary Studies	**0.33**	**0.54**	**0.69**	**0.90**	**Average**
15	0.17	0.11	0.07	0.02	0.09
50	0.09	0.06	0.04	0.01	0.05
100	0.07	0.04	0.03	0.01	0.03
150	0.05	0.03	0.02	0.01	0.03
Average	0.10	0.06	0.04	0.01	0.05

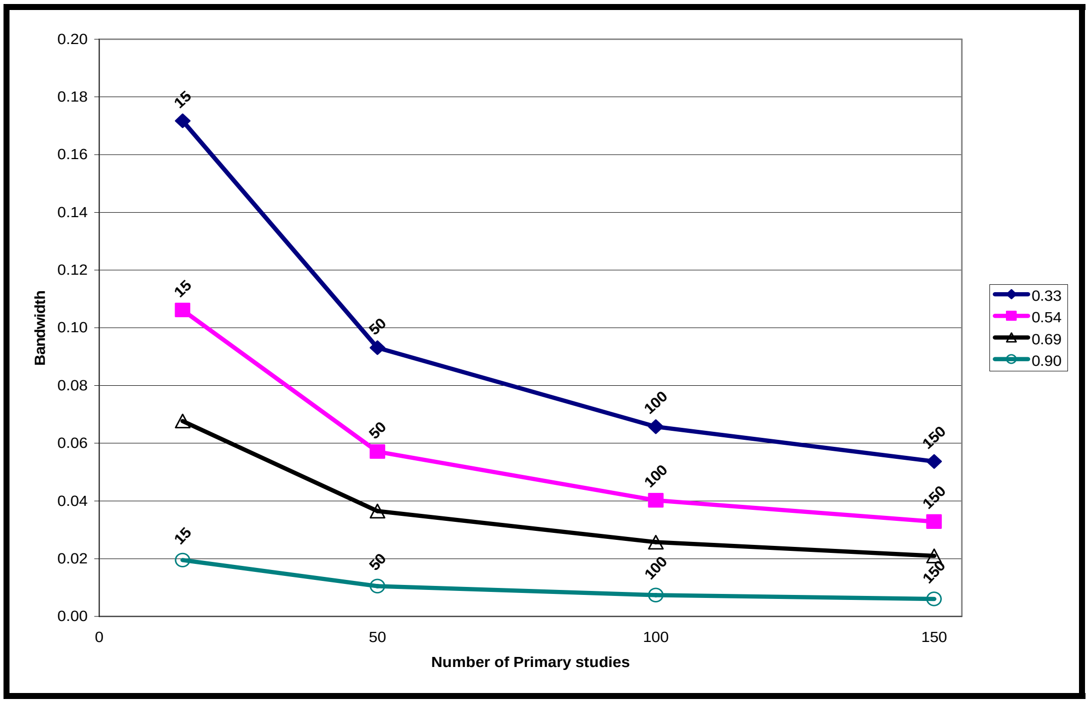


Figure 17. Band width of estimated mean reliability for number of primary studies by coefficient alpha

A Deeper Look at Band Coverage

When the results for average *Band Coverage* were first examined there were a noticeable number of the simulations where the average *Band Coverage* was quite small. Recall that for each condition investigated, several RG analyses were simulated such that the value of the average *Band Coverage* is the average proportion of times (for the 1,000 to 10,000 replications that were simulated) that the actual population parameter was within a 95% confidence band around the mean reliability estimate's value. In other words, if the average *Band Coverage* for a particular estimate was .30, this means that for the RG analysis for that particular set of factors 30% of the confidence bands contained the population coefficient alpha and 70% of them did not. As a means to evaluate these results the *Band Coverage* was divided into three categories such that *Band Coverage* that was less than .50 was considered "small" and *Band Coverage* greater than or equal to .50 and less than .925 was considered "medium" and *Band Coverage* greater than or equal to .925 and less than or equal to 1 was considered "large." The "cut off" values chosen for "large" were based on Bradley's (1978) approach to defining robustness. Using these categories, approximately 34.8% of all 6,400 conditions had a small average *Band Coverage,* approximately 34.47% had medium average *Band Coverage,* and 30.73% had large *Band Coverage*.

In addition to investigating the overall proportions of small, medium and large coverage, the extent to which the type of treatment by each factor resulted in robust (i.e., large) *Band* Coverage also was analyzed. Table 27 displays the results for type of treatment. These values in the table represent the percentage of the total number of conditions run for each type of treatment. For example, in Table 27, the value 11.72%

appears in the cell that is the intersection of *Violation* and large *Band Coverage*. This value represents the proportion of all of the 1,280 conditions that were simulated for the treatment *Violation* that had *Band Coverage* that was greater than or equal to .925 and less than 1. When the treatment was *HLM* about 23.28% of the *Band Coverage* was large. When the treatment was *Random* only about 21.48% was large. The treatments *Mean* and *Median* had the largest percentage of *Band Coverage* that was large, 51.95% and 45.23%, respectively.

Table 27

Large Band Coverage by Type of Treatment

	Treatment					
Band Coverage	**Violation**	**Random**	**Median**	**Mean**	**HLM**	**Total**
Large	11.72%	21.48%	45.23%	51.95%	23.28%	30.73%
Total	1280	1280	1280	1280	1280	6400

Next, the percentage of large *Band Coverage* for factors by treatment is presented. Table 28 displays the results for intra-class correlation by treatment. The values in each cell represent the percent of large *Band Coverage* for all the conditions simulated that shared those characteristics. For example, in Table 28, the cell where ICC = 0 and the treatment is *Violation* contains the value 18.44%. In this study there were 320 conditions simulated for each value of ICC and treatment. The 18.44% represents the percentage of those 320 conditions that were simulated such that ICC = 0 and the treatment was *Violation.* For each value of ICC there were 1,600 conditions that were simulated. The percentage for each row in the last column represents the percentage of the 1,600 conditions where the ICC had large *Band Coverage*. For example, 43.19% of the 1,600 conditions simulated where ICC = 0 had large *Band Coverage*. Notice in this table that as the intra-class correlation increased the percentage of large *Band Coverage* decreased such that when ICC = .90 only 7.69% of the 1,600 conditions had large *Band Coverage.* Out of the 320 simulations each for *Violation* and *HLM* only 1.25% and 1.56%, respectively, had large *Band Coverage* when ICC = .90. While the *Mean* treatment seemed to have the highest percentage of large *Band Coverage* it was still only 69.69% when ICC = 0 and was as small as 17.50% when ICC= .90. Probably the most disconcerting result was the fact that even when ICC = 0 (i.e., no proportion of variance in reliability that is between studies), only 43.19% of the 1,600 conditions simulated had large *Band Coverage.*

Table 28

Percentage of Large Band Coverage for Intra-class Correlation by Treatment

			Treatment			
ICC	**Violation**	**Random**	**Median**	**Mean**	**HLM**	**Total**
0.00	18.44%	33.44%	61.25%	69.69%	33.13%	43.19%
0.01	19.06%	31.25%	59.69%	65.63%	34.38%	42.00%
0.30	8.13%	15.94%	47.19%	55.00%	24.06%	30.06%
0.90	1.25%	5.31%	12.81%	17.50%	1.56%	7.69%
Total	11.72%	21.48%	45.23%	51.95%	23.28%	30.73%

The results for percentage of large *Band Coverage* for coefficient alpha by treatment are presented in Table 29. As with the results for the intra-class correlation and treatment, the *Mean* treatment had the highest percentage of large *Band Coverage* for each value of coefficient alpha. When the treatment was *HLM* and ρ_{xx} = .33, 55.00% of the 320 conditions that were simulated had large *Band Coverage*; this was more than three times as much as when the treatment was *Violation.* Notice that out of the 320 conditions that were simulated such that ρ_{xx} = .54 and the treatment was *Violation*, only 0.94% had large *Band Coverage.* In general, when ρ_{xx} = .33 or .69 the percentage of large *Band Coverage* was almost twice as much as when ρ_{xx} = .54 or .90. In addition, when ρ_{xx} = .90 and the treatment was *HLM* only 7.19% of the 320 conditions simulated had large band coverage. Not only was this a smaller value than when the treatment was *Violation*, it was the smallest percentage of large *Band Coverage* when $\rho_{xx} = .90$. Though $\rho_{xx} = .90$ is usually considered a desirable value for coefficient alpha, in this study only 22.00% of the 1,600 conditions that were simulated had *Band Coverage* that would be considered robust.

Table 29

Percentage of Large Band Coverage for Coefficient Alpha by Treatment

			Treatment			
ρ_{xx}	**Violation**	**Random**	**Median**	**Mean**	**HLM**	**Total**
0.33	16.88%	20.31%	45.94%	65.00%	55.00%	40.63%
0.54	0.94%	4.38%	41.56%	47.50%	9.38%	20.75%
0.69	21.56%	40.00%	58.13%	56.56%	21.56%	39.56%
0.90	7.50%	21.25%	35.31%	38.75%	7.19%	22.00%
Total	11.72%	21.48%	45.23%	51.95%	23.28%	30.73%

The results for percentage of large *Band Coverage* for average sample size by treatment are presented in Table 30. Once again, the *Mean* treatment had the highest percentage of *Band Coverage* for each of the average sample size values. However, when the average sample size was 500, the percentage of large *Band Coverage* for the *Mean* treatment was only slightly larger than that for the *Median* treatment (41.25% and 40.94%, respectively). When the average sample size was equal to 10 the percentage of large *Band Coverage* for the *HLM* treatment was slightly less than the percentage for *Violation* and more than half that for *Random.* However, for the larger values of average sample size the *HLM* treatment had a higher percentage of large *Band Coverage* than *Violation* and *Random*. When the average sample size was 500, the percentage of large *Band Coverage* for *HLM* was almost three times as much as that for *Random* and more than four times as much as that for *Violation*. In general, the smaller average sample sizes had a higher percentage of large *Band Coverage* than did the larger average sample size. When the average sample size was 10, the percentage of large *Band Coverage* was still only 40.94% and was only 26.44% when the average sample size was 500.

Table 30

Percentage of Large Band Coverage for Average Sample Size by Treatment

	Treatment					
Average Sample Size	**Violation**	**Random**	**Median**	**Mean**	**HLM**	**Total**
10	24.69%	43.44%	50.94%	66.56%	19.06%	40.94%
50	5.63%	16.56%	44.69%	51.88%	21.88%	28.13%
100	9.06%	14.06%	44.38%	48.13%	21.56%	27.44%
500	7.50%	11.88%	40.94%	41.25%	30.63%	26.44%
Total	11.72%	21.48%	45.23%	51.95%	23.28%	30.73%

The results for percentage of large *Band Coverage* for the number of primary studies by treatment are presented in Table 31. As with results for the other factors presented thus far, the *Mean* treatment had the highest percentage of large *Band Coverage* for each of the primary studies values. Under these conditions the *Median* treatment had parentages that were not much smaller than when the *Mean* treatment was used. Also, the treatment *HLM* had percentages that were always slightly higher than those for the *Random* and almost twice as much as when the treatment was *Violation*. In general, the larger the number of primary studies the lower the percentage of large *Band Coverage*. When the number of primary studies was 150 only 19.94% of the 1,600 conditions that were simulated had large *Band Coverage* and when the number of primary studies was 15 only 46.75% of the 1,600 conditions had large *Band Coverage.* The results for the percentage of large *Band Coverage* for the number of reliabilities for each primary study by treatment are presented in Table 32. Because the number of reliabilities had five possible values the number of conditions generated for each cell was 256 not 320, (e.g., there were 256 conditions simulated such that the number of reliabilities was 1 and the treatment was *Violation*). In addition, there were 1,280 conditions simulated, not 1,600, for each number of reliabilities value. When the number of reliabilities was 1 and the treatment was *HLM*, the percentage of large *Band Coverage* was 30.47%; for the other treatments it was 21.88%. As the number of reliabilities increased, the percentage of large *Band Coverage* increased for the *Mean* treatment. When the treatment was *Median* and the number of reliabilities was 2, the percentage of large *Band Coverage* was more than double the percentage when the number of reliabilities was 1. However, the percentage of large *Band Coverage* was slightly smaller for the *Median* treatment when the number of reliabilities was 3 and then increased when the number of reliabilities was 10 and then 50. When the treatment was *HLM* and the number of reliabilities was 2 the percentage of large *Band Coverage* was only slightly larger than when the number of reliabilities was 1. When the treatment was *Random* the percentage of large *Band Coverage* decreased as the number of reliabilities increased except when the number of reliabilities increased from 10 to 50. In this case, the percentage of large *Band Coverage* went from 19.92% to 21.88%, respectively. When the number of reliabilities was 3 and the treatment was *HLM* the percentage of large *Band Coverage* decreased to 28.52% and was even smaller when the number of reliabilities was increased. When the number of reliabilities was 50 and the treatment was *HLM* only 8.20% of the 256 conditions simulated had large *Band Coverage*. Finally, when the treatment was *Violation* the percentage of large *Band Coverage* decreased as the number of reliabilities increased. When the number of reliabilities was 50 and the treatment was *Violation* only 2.73% of the 256 conditions simulated had large *Band Coverage*. Overall,

when the number of reliabilities was 1, only 23.59% of the 1,280 conditions simulated had large *Band Coverage*. For the other number of reliabilities values the percentage of large *Band Coverage* was about one third of the 1,280 conditions simulated for each reliability value.

Table 31

Percentage of Large Band Coverage for the Number of Primary Studies by Treatment

	Treatment					
Number of Primary Studies	**Violation**	**Random**	**Median**	**Mean**	**HLM**	**Total**
15	19.06%	37.50%	66.88%	69.69%	40.63%	46.75%
50	11.56%	19.38%	47.81%	55.94%	22.50%	31.44%
100	9.06%	17.81%	35.63%	45.31%	16.25%	24.81%
150	7.19%	11.25%	30.63%	36.88%	13.75%	19.94%
Total	11.72%	21.48%	45.23%	51.95%	23.28%	30.73%

Table 32

Percentage of Large Band Coverage for Number of Reliabilities Per Study by Treatment

	Treatment					
Number of Reliabilities	**Violation**	**Random**	**Median**	**Mean**	**HLM**	**Total**
1	21.88%	21.88%	21.88%	21.88%	30.47%	23.59%
2	14.45%	22.27%	48.44%	48.44%	32.03%	33.13%
3	10.94%	21.48%	42.58%	56.64%	28.52%	32.03%
10	8.59%	19.92%	54.69%	64.84%	17.19%	33.05%
50	2.73%	21.88%	58.59%	67.97%	8.20%	31.88%
Total	11.72%	21.48%	45.23%	51.95%	23.28%	30.73%

Summary of Results

These results were evaluated by first looking at the five choices of treatments (*Violation, Mean, Median, Random,* and *HLM*) addressed in the research questions for this study. This was executed by creating box plots for these treatments and each of the outcomes: *Bias, RMSE, Band Coverage* and *Band Width.* The box plots indicated that regardless of the type of treatment used, the results were about the same for *Bias, RMSE,* and *Band Width*; that is, treatment does not impact the variability in these three outcome variables. However, the type of research design did appear to have an impact on the variability in *Band Coverage*. This was confirmed when eta-squared was calculated using PROC GLM in SAS such that the dependent variables were *Bias, Root Mean Square Error, Band Coverage*, and *Band Width* and the independent variables were the five types of factors (magnitude of coefficient alpha, average sample size, number of journal studies, number of reliability coefficients from each journal study, and the magnitude of the intra-class correlation), and the choice of treatment. For all four of the dependent variables, eta-squared was calculated for the main effect along with the first-order interactions of the independent variables. Even though the choice of treatment was only a significant main effect for *Band Coverage*, the impact of treatment was included in the results for the evaluation of all of the significant main effects and first-order interactions for *Bias, Root Mean Square Error, Band Coverage*, and *Band Width* because their impact was the main focus of this research study.

Even though the eta-squared results indicated that ICC, ρ_{xx}, and *n,* in addition to the interaction between ICC and $\rho_{xx,}$ and the interaction between treatment and ρ_{xx} all had an impact on the variability in *Bias*, overall, the magnitude of the *Bias* was not large, and in all cases the estimated mean reliability was never overestimated. While these results did indicate that the *Median* treatment resulted in slightly larger values of *Bias*, these values never exceeded .05 in magnitude. These results for *Bias* suggest that regardless of the treatment or the other factors investigated in this simulation, the estimated mean reliability was not overestimated and was only slightly underestimated.

The main effect factors that had a significant impact on the variability in *RMSE* were $\rho_{xx,}$ n, and ICC. In addition the interaction between ρ_{xx} and n, the interaction between ρ_{xx} and ICC, and the interaction between ρ_{xx} and the type of treatment also were shown to have a significant impact on the variability in *RMSE*. These results suggest that smaller values of ρ_{xx} will have a slightly larger *RMSE* compared to larger values of ρ_{xx} and larger samples sizes have a slightly smaller and somewhat more stable *RMSE* than the smaller sample sizes. Even though when ICC = .90 the average RMSE was slightly larger than the results for the smaller values of ICC, overall the magnitude of ICC did not appear to have a large impact on the variability in *RMSE.* Overall, the *RMSE* was never very large, which would suggest that the reliability estimates were somewhat stable regardless of the magnitude of any of these factors.

When *Band Coverage* was examined, the main effect factors that had a significant impact on the variability in *Band Coverage* were ICC and the number of primary studies. In addition, the choice of treatment also had a significant impact on the variability in *Band Coverage*. The first-order interactions that were significant were the interaction between ICC and $\rho_{xx,}$ and the interaction between ρ_{xx} and $n.$ These results suggest that larger values of ICC will have a much smaller *Band Coverage* compared to smaller values of ICC. More important, the type of treatment does not seem to improve the size of the *Band Coverage* as ICC increases. However, these results also suggest that there is a noticeable interaction between ICC and ρ_{xx} when it comes to *Band Coverage.* When ρ_{xx} = .33, .54, or .69, the band coverage was somewhat the same and consistent in behavior (i.e., the smaller the ICC the larger the *Band Coverage*); this was not the case for ρ_{xx} =.90. The average *Band Coverage* for ρ_{xx} =.90 was very small, .47, and the larger ICC resulted in an increase in *Band Coverage*. A similar pattern was seen for the interaction between ρ_{xx} and average sample size. As sample size increased the average *Band Coverage* for ρ_{xx} = .33, .54, and .69 was relatively stable. For ρ_{xx} = .90, the average *Band Coverage* for *n* = 50, 100, and 500 went from .60 to .33 to .06, respectively.

The main effects factors that had an impact on the variability in *Band Width* were *n*, $\rho_{xx,}$ and the number of primary studies. Ironically, the first-order interaction effects that were significant were all some pairing of these three factors: the interaction between ρ_{xx} and *n*, the interaction between *n* and NPS, and the interaction between ρ_{xx} and NPS. In contrast to the other three dependent variables, *Bias, RMSE,* and *Band Coverage*, the η^2 value for ICC was approximately 0. The variability in ICC did not have much of an impact on the variability in *Band Width.* These results suggest an inverse relationship between average sample size and *Band Width* such that the larger the average sample size, the smaller the *Band Width.* This was also the case when evaluating the impact of the magnitude of the population reliability parameter and the number of primary studies. The larger the magnitude of ρ_{xx} or the larger the number of primary studies, the smaller the *Band Width.* There was also a noticeable interaction between ρ_{xx} and average sample size for *Band Width* such that when the average sample size is small and ρ_{xx} is small, there was much more variability in *Band Width* than when average sample size was large and ρ_{xx}. Similar results were seen for the interaction for these two factors, *n* and ρ_{xx}, with the number of primary studies. In general, smaller values produced wider bands.

Overall, the *Band Width,* regardless of the factors, was quite small. *Band Width* never exceeded .27 and on average was .05. The confidence bands were very narrow.

Because there was a noticeable number of simulations in which the average *Band Coverage* was quite small (i.e., less than .50), the *Band Coverage* was evaluated by dividing the results into three categories such that the *Band Coverage* was considered small if it was greater than zero but less than .50; *Band Coverage* greater than or equal to .50 and less than .925 was considered "medium," and *Band Coverage* greater than or equal to .925 and less than or equal to 1 was considered "large." The "cut off" values chosen for "large" were based on Bradley's (1978) approach to defining robustness such that the percentage of large *Band Coverage* would be those conditions whose results were fairly robust to Type I Error. Using these categories, approximately 34.8% of all 6400 conditions had a small average *Band Coverage,* approximately 34.47 % had medium average *Band Coverage* and 30.73% had large *Band Coverage.*

In addition to examining the overall percentage of small, medium, and large *Band Coverage*, the parentage of conditions that resulted in *Band Coverage* for each treatment and for each factor by treatment also was evaluated. Out of the 1,280 conditions simulated for each treatment, the *Mean* treatment had the highest percentage of large *Band Coverage* (51.95%), and the lowest percentage was for the *Violation* treatment (11.72%). When these treatments were paired with the factors investigated in this study (i.e., intra-class correlation, coefficient alpha, average sample size, number of primary studies, and number of reliabilities per study), the *Mean* treatment usually had the highest percentage of large *Band Coverage*. This was true for every factor regardless of the value with two exceptions. First, when ρ_{xx} = .69, the *Median* treatment had a slightly higher large *Band Coverage* than the *Mean* treatment (58.13% and 56.56%, respectively). Second, when the number of reliabilities per study was equal to 1, the *HLM* treatment had the highest percentage of large *Band Coverage* out of the five treatments and the other four treatments had band coverage that were all equal to 21.88%. In general, however, the *HLM* treatment usually had very small percentage of large *Band Coverage* and in many instances the results were very similar to the results when the treatment was *Violation*.

For each of the five factors evaluated even when the value for each was at a "desirable" level, the percentage of large *Band Coverage* was remarkably small. For example, even when ICC = 0 only 43.19% of the 1,600 conditions simulated had large *Band Coverage*. When ρ_{xx} = .90, only 22.00 % of the 1,600 conditions simulated had large *Band Coverage*. When the average sample size was equal to 500, only 26.44% of the conditions simulated had large *Band Coverage*. When the number of primary studies was equal to 150 only 19.94% of the conditions had large *Band Coverage*. Finally, when the number of reliabilities per study was 1 (a somewhat desirable number) the percentage of large *Band Coverage* was only 23.59%. One might debate what values for the factors examined would be considered "desirable" for an RG meta-analysis. However, it is quite obvious that when only 30.73% of the conditions simulated had *Band Coverage* that was robust none of the values for the factors and none of the treatments really had very "desirable" results.

Chapter Five:

Conclusions

Summary of the Study

Both validity and reliability indices are a function of the scores on a measure, and the magnitude of these indices can fluctuate across administrations of a measure. It is a common mistake to say that a test is reliable when in fact it is not the test that is reliable, but the scores on a test that are reliable. Because reliability can fluctuate across studies, it has been recommended that researchers should always evaluate the reliability of their measure and report the results (Wilkinson & APA Task Force on Statistical Inference, 1999). In 1998, Vacha-Haase addressed this issue when she proposed a fixed-effects meta-analytic method for evaluating reliability, similar to validity generalization studies, called reliability generalization (RG). Validity generalization studies have been conducted to describe the extent to which validity evidence for scores are generalizable across research contexts (Hunter & Schmidt, 1990; Schmidt & Hunter, 1977). In a similar fashion, RG studies can be used to investigate the distribution of reliability estimates across studies and to identify study characteristics that may be related to variation in reliability estimates, such as sample size, type of reliability estimate (coefficient alpha vs. test-retest), different forms of an instrument, or participant characteristics (Henson, 2001; Vacha-Haase, 1998). This method is recommended for describing estimated measurement error in a test scores across studies and also can be used to analyze measurement error in differences.

This research primarily focused on appropriate analysis of reliability estimates that are not statistically independent. The assumption of independence of observations is commonly violated in meta-analytic research (Becker, 2000; Hedges & Olkin, 1985; Hunter & Schmidt, 1990). As the available literature suggests, violating the assumption of independence is a serious issue.

There are several approaches to dealing with the violation of independence that have been recommended by researchers (see Becker, 2000). Some of these are, ignoring it and treating each observation as independent (e.g., Smith, Glass, & Miller, 1980), calculating one mean or median from each study (e.g., Tracz et al., 1992), selecting only one observation per study (e.g., Rosenthal & Rubin, 1986), or using a mixed effects model (e.g., Beretevas & Pastor, 2003).

For this study the samples of primary studies were generated using population parameters from a three-parameter IRT model (Table 2) that was developed by Hanson and Beguin (1999). From these simulated examinee responses, subsets of items were selected that yielded the target values of coefficient alpha. These target values, computed from the simulated examinees, were used as the population values to which the subsequent sample estimates were compared. The coefficient alpha values were generated using the information from the three-parameter model using the item information from the ACT Mathematics Assessment.

The research was conducted using a Monte Carlo simulation study method in which random samples were simulated under known and controlled population conditions. In the Monte Carlo study, RG studies were simulated by generating samples in primary studies, estimating reliability of scores in these samples, and then aggregating the sample reliability estimates in the RG studies. The Monte Carlo study included five factors. These factors were (a) the magnitude of the coefficient alpha (with $\rho_{xx} = 0.30$, 0.50, 0.70, and 0.90), (b) sample size in the primary studies (average sample sizes, *n*, of 10, 50, 100, and 500), (c) number of primary studies in the RG study (with $k = 15$, 50, 100, and 150), (d) number of reliability estimates from each study (with $i = 1$, 2, 3, 10 and 50) and (e) the degree of violation of independence where the strength of the dependence is related to the number of reliability indices (i.e., coefficient alpha) derived from a simulated set of examines and the magnitude of the correlation between the journal studies (intra-class correlation $\rho = 0$, .0l , .30, and .90). The values chosen for each of these factors were based in part on observed factors of actual RG studies, in part on factors of the Tracz et al. (1992) simulation study, and mostly on values that represent a range that is reasonable and typical in simulation studies.

In addition, there were five types of treatments that were applied: first, the dependence was ignored, referred to as *Violation*. Second, a reliability index was randomly selected from each of the simulated journal studies, referred to as *Random*. Third, a mean was calculated from each simulated journal study, referred to as *Mean*. Fourth, a median was calculated from each simulated journal study, referred to as *Median.* Finally, a two-level mixed model was used to calculate the estimated mean reliability using a null model such that the intercept value was the average reliability; this was referred to as *HLM.*

The research was conducted using SAS/IML version 9.1. Conditions for the study were run under Windows XP. Normally distributed random variables were generated using the RANNOR random number generator in SAS. A different seed value for the random number generator was used in each execution of the program and the program code was verified by hand-checking results from benchmark datasets. The SAS PROC MIXED procedure was used to generate the two level null models used in this study.

The impact of the treatment factors was evaluated based upon the bias in the mean estimates, root mean square estimates, the confidence band coverage, and the average confidence band width.

Research Questions

In RG studies the dependent variable in the analyses is the reliability estimate (Henson & Thompson, 2001). This research focused on how certain study methods, in regards to violation of independence, affect the estimated mean reliability of scores calculated across studies. The key questions that were addressed in this study were:

1. What is the effect on point and interval estimates of mean reliability of ignoring violation of independence of observations in RG studies (i.e., treating all reliability coefficients as independent observations)?
2. What is the effect on point and interval estimates of mean reliability of using a mean or median reliability from each study as part of a sample in a RG study?
3. What is the effect on point and interval estimates of mean reliability of randomly selecting a reliability estimate from each study as a part of a sample in a RG study?
4. What is the effect on point and interval estimates of mean reliability of using a two level mixed-effects model for RG studies (i.e., reliabilities are nested within studies)?
5. In regard to violations of independence, what impact do factors such as the magnitude of coefficient alpha, sample size, number of journal studies, number of reliability coefficients from each study, and the magnitude of the intra-class correlation (ICC) of

the studies (i.e., the magnitude of the violation of independence) have when any of the methods discussed in the four research questions above are investigated?

Summary of Study Results

Because the first four research questions addressed the impact of the type of treatment, these results were evaluated by first looking at the impact of the treatments used in this study (*Violation, Mean, Median, Random,* and *HLM*). This was carried out by creating box plots for these treatments and each of the outcomes, *Bias, RMSE, Band Coverage* and *Band Width.* The box plots indicated that the types of treatment does not impact the variability in *Bias, RMSE,* and *Band Width* but did seem to have an impact on *Band Coverage.* This was later confirmed when eta-squared was calculated in regards to the type of treatment and their interaction with the other factors investigated in this study.

Eta-squared was calculated using PROC GLM in SAS such that the dependent variables were *Bias, RMSE, Band Coverage*, and *Band Width* and the independent variables were the five types of factors (magnitude of coefficient alpha, average sample size, number of journal studies, number of reliability coefficients from each journal study, and the magnitude of the intra-class correlation), and the choice of treatment. For all four of the dependent variables eta-squared was calculated for the main effects and first-order interactions of the independent variables. A cut off value of $\eta^2 \geq 0.05$ was used to determine which factors had an important impact on the dependent variables. While the type of treatment was only a main effect for *Band Coverage*, because the impact of treatment is addressed in the research questions in this study, treatment included in all of the results for the evaluation of *Bias, RMSE, Band Coverage*, and *Band Width.*

Eta-squared results indicated that ICC, ρ_{xx}, and *n,* in addition to the interaction between ICC and $\rho_{xx,}$ and the interaction treatment and ρ_{xx} all had an impact on the variability in *Bias*; overall, the magnitude of the *Bias* was not large and in all cases the estimated mean reliability was never overestimated.

The main effect factors that had an impact on the variability in *RMSE* were $\rho_{xx,}$ *n*, and ICC. In addition, the interaction between ρ_{xx} and *n*, the interaction between ρ_{xx} and ICC, and the interaction between ρ_{xx} and the type of treatment, also were shown to have an impact on the variability in *RMSE*. These results suggest that smaller values of ρ_{xx} had a slightly larger *RMSE* compared to larger values of ρ_{xx} and larger samples sizes have a slightly smaller and somewhat more stable *RMSE* than the smaller sample sizes. Even though when ICC = .90 the average RMSE was slightly larger than were the results for the smaller values of ICC, one could argue that ICC was not that influential on *RMSE.* Overall, the *RMSE* was never very large, which would suggest that the reliability estimates were somewhat stable regardless of the magnitude of any of these factors.

When *Band Coverage* was examined, the main effect factors that showed a significant impact on the variability in *Band Coverage* were ICC, and the number of primary studies. In addition, the type of treatment also had a significant impact on the variability in *Band Coverage*. The first-order interactions that were significant were the interaction between ICC and $\rho_{xx,}$ and the interaction between ρ_{xx} and *n.* These results suggest that larger values of ICC had a much smaller *Band Coverage* compared to smaller values of ICC. In addition, the type of treatment did not seem to improve the size of the *Band Coverage* as ICC increases. However, these results also suggest that there is a noticeable interaction between ICC and ρ_{xx} when it comes to *Band Coverage*. A similar pattern was seen for the interaction between ρ_{xx} and average sample size. As sample size increased the average *Band Coverage* for $\rho_{xx} = .33$, but when $\rho_{xx} = .90$ the average *Band Coverage* had much more variability and decreased substantially as the average sample size increased.

The main effects factors that had an impact on the variability in *Band Width* were *n*, $\rho_{xx,}$ and the number of primary studies. The first-order interaction effects that were significant were all some pairing of these three factors: the interaction between ρ_{xx} and *n*, the interaction between *n* and NPS, and the interaction between ρ_{xx} and NPS. In contrast to the other three dependent variables, *Bias, RMSE,* and *Band Coverage*, the variability in ICC did not have much of an impact on the variability in *Band Width.* These results suggest that there was an inverse relationship between average sample size and *Band Width* such that the larger the average sample size the smaller the *Band Width.* This was also the case when evaluating the impact of the magnitude of the population reliability parameter and the number of primary studies. The larger the magnitude of ρ_{xx} or the larger the number of primary studies the smaller the *Band Width.* There was also a noticeable interaction between ρ_{xx} and average sample size for *Band Width* such that when the average sample size is small and ρ_{xx} is small there was much more variability in *Band Width* then when average sample size was large and ρ_{xx}. Similar results were seen for the interaction for these two factors, *n* and ρ_{xx}, with the number of primary studies. In general, smaller values produced wider bands. Overall the *Band Width,* regardless of the factors, was quite small. *Band Width* never exceeded .27 and on average was .05. The confidence bands were very narrow.

Because there was a noticeable number of simulations where the average *Band Coverage* was quite small (i.e., less than .50), the *Band Coverage* was evaluated by dividing the results into three categories such that the *Band Coverage* was considered small if it was greater than zero but less than .50; *Band Coverage* greater than or equal to .50 and less than .925 was considered "medium," and *Band Coverage* greater than or equal to .925 and less than or equal to 1 was considered "large." The "cut off" values chosen for "large" were based on Bradley's (1978) approach to defining robustness such that the percentage of large *Band Coverage* would be those conditions whose results were fairly robust to Type I Error. Using these categories, approximately 34.8% of all 6,400 conditions had a small average *Band Coverage,* approximately 34.47 % had medium average *Band Coverage,* and 30.73% had large *Band Coverage*. In regards to the different types of treatments, the *Mean* research had the largest percentage of *Band Coverage* that was robust (51.95%).

When the treatments were paired with the factors investigated in this study the *Mean* treatment usually still had the highest percentage of large *Band Coverage*. This was true for every factor regardless of the value with two exceptions. First, when ρ_{xx} = .69, the *Median* treatment had a slightly higher large *Band Coverage* than the *Mean* treatment (58.13% and 56.56%, respectively). Second, when the number of reliabilities per study was equal to 1, the *HLM* treatment had the highest percentage of large *Band Coverage* out of the five treatments and the other four treatments had band coverage that were all equal to 21.88%. In general, however, the *HLM* treatment usually had very small percentage of large *Band Coverage* and in many instances the results were very similar to the results when the treatment was *Violation*.

For each of the five factors evaluated, even when the value for each was at a "desirable" value, the percentage of large *Band Coverage* was still very small. For example, even when ICC = 0 only 43.19% of the 1,600 conditions simulated had large *Band Coverage*. When $\rho_{xx} = .90$, only 22.00 % of the 1,600 conditions simulated had large *Band Coverage*. When the average sample size was equal to 500 only 26.44% of the conditions simulated had large *Band Coverage*. When the number of primary studies was equal to 150 only 19.94% of the conditions had large *Band Coverage*. Finally, when the number of reliabilities per study was 1 (a somewhat desirable number) the percentage of large *Band Coverage* was only 23.59%. In general because only 30.73% of all the conditions simulated had *Band Coverage* that was robust it could be argued that most of the values for the factors and most of the treatments did not have very "desirable" results.

Discussion

It was expected, based on previous research (Beretevas & Pastor, 2003), that *HLM* would provide better point estimates and better interval estimates than the rest of the treatments applied; however, this was not the case with this study. When the type of treatment was investigated as a part of the other factors, at times *HLM* behaved more like *Violation* than any of the other treatments. This could be because of the five types of treatments investigated these two methods were the only two that used all the observations as a part of estimating the mean reliability. In general, there was very little *Bias* in the results and the *RMSE* results were relatively small. In addition, the *Band*

Width was overall very small which would explain the overall poor *Band Coverage*, i.e., narrow bands would "capture" fewer estimates. When the *Band Coverage* was evaluated only 30.73%, or less than one third of all the conditions simulated, had *Band Coverage* that was considered robust. The fact that *HLM* and *Violation* both had results where the percentage of large *Band Coverage* was very small would indicate that these two types of treatments are likely to produce reliability that are less likely to fall within a 95% confidence interval. Based on these results calculating a mean from each study seemed to produce the most robust *Band Coverage.* Even though it was better, the average *Band Coverage* for this type of treatment was still only .72.

These results did suggest that the magnitude of ICC, the magnitude of the population reliability parameter, and the magnitude of the average sample size do have an impact on the point and interval estimates results. The number of primary studies had some impact in regards to *Band Width* but the number of reliabilities from each study was not seen to be a contributing factor. Based on these results it could be argued that the point and interval estimates are impacted the most when ICC, the population reliability parameter, and the average sample size are rather large. As was seen in these results, when this occurred the *Band Coverage* was quite small. However, for *Bias* and *RMSE*, even though significant differences in the variability of the factors were found, overall, these values were rather small. Another value that was consistently small was the *Band Width.* These results suggest that while factors like population reliability parameter and the average sample size do have an impact on the variability of the outcomes, the overall averages for these outcomes was rather small. However, the magnitude of the ICC alone and its interaction with these factors can have an impact on the point and interval estimates of reliability.

Limitations of the Study

The limitations of this study are related to the Monte Carlo method for the study. While the Monte Carlo method was used to simulate RG studies, the values of the factors used in the simulation were fixed for each study. Specifically, because the data for this study were simulated the number of reliability indices from each simulated study was a fixed value in each of the simulations (i.e., each study contributed the same number of reliability indices per study). While it is obvious that several of the RG studies conducted so far are treating reliability coefficients from the same study as independent, it is also obvious that not all of the studies contribute equal amounts of reliability coefficients. In addition, because the models in this study are fixed-effects models small sample sizes should not be a concern (Randenbush & Bryk, 2002).

In several of the RG studies conducted so far test-retest reliability estimates given are very rare and seldom evaluated. Because coefficient alpha is the most common reliability coefficient reported, this was the only index used in the study. It is important to note however, that coefficient alpha has a tendency to underestimate the actual reliability index (Crocker & Algina, 1986).

The data for this study were simulated using information from a test of ability. All of the RG studies that been conducted thus far have investigated reliability in the context of an instrument that measures some type of psychological construct. Measures of ability have a tendency to have more variability than measures of psychological constructs. It is possible that in the actual RG studies there may have been less variability in the results of the measures investigated. This difference in variability could have an impact on the mean estimates of the reliability indices.

Another possible limitation to this study is the fact that in each RG analysis estimates were investigated using z transformation for coefficient alpha,

$$z = \ln\left(1 - |\alpha|\right)$$

to normalize the sampling distributions where the transformed value of coefficient alpha is approximately normally distributed with a variance of $k/\{2(k\text{-}1)(N\text{-}2)\}$(Bonett, 2002). According to Felt and Charter (2006), there are many ways of averaging reliability across studies; perhaps another method may have led to different results.

Implications

Importance of the Study. Researchers have suggested that the use of *HLM* should provide a better model to investigate the variability of score reliabilities (Beretevas et al., 2002; Beretevas & Pastor 2003). The results of this study suggest that while the magnitude of the intra-class correlation has a significant impact on the variability in the *Bias* and the *RMSE,* the impact on both of these independent variables is negligible. When independence is violated, the point estimates are still relatively stable and only slightly underestimated, regardless of the type of treatment that might be used to estimate the mean coefficient alpha. This research does indicate that researchers should be careful in regards to constructing confidence intervals because the *Band Width* was on average .06, the average *Band Coverage* was only .61, and only 30.73 % of the simulations had coverage that was robust.

Importance in Regards to Future RG Studies. While this research seems to indicate that using HLM is not necessarily the best solution for controlling for dependence, it is possible that the use of mixed models may provide more power in RG analyses such that this method may provide more control of Type II error for testing differences in group means. More research needs to be done in this area to investigate the impact this method may have on tests of group differences.

Because the magnitude of ICC does seem to impact the stability of the results, future RG studies should consider the magnitude of the ICC that might be present. While it is usually not possible to calculate the population parameter for this index, one could still estimate it from the sample. Regardless of the type of treatment employed, this research still supported the assumption that the larger the ICC the more problematic the results.

While these results indicated that the point estimates calculated from a RG analysis have very little bias regardless of the magnitude of the factors or the type of treatment, the RG researcher should probably not use these point estimates to build confidence intervals for inferential statistics. As Felt and Charter (2006) point out, the average reliability obtained from averaging across studies is not the same as the average that would be obtained if all of the raw data from the groups of interest were available and the researcher calculated coefficient alpha from the combination of the groups. They argue that the average reliability obtained from averaging across studies should never be the value used to construct confidence intervals for tests of significance among coefficients. They do suggest that there are methods for combining coefficient alpha across that would produce the same value as if one did have all the raw data (see Charter 2003).These results suggest that RG meta-analysis may be useful in estimating what is typical reliability for a given measure or construct but should not be used when creating confidence bands.

Suggestions for Future Research

While this study did explore several factors and different types of treatments in regards to reliability generalizations studies, this research only explored the average reliability across studies without considering possible moderating variables. This research should be reproduced such that values of common moderating variables that are present in typical RG studies can be explored such as sample size, different forms of an instrument, or participant characteristics (Henson, 2001; Vacha-Haase, 1998).

Because the data for this analysis were generated from a dichotomously scored measure of mathematics ability, this research should be replicated using simulated data from a measure of a psychological construct. In addition, this study also could be duplicated using actual data from an RG study where moderating variables and a measure of a psychological construct were evaluated. Also this research did not consider the issues in regards to reliability in longitudinal studies. It is possible that longitudinal studies will produce rather large intra-class correlation (DeShon, Ployhart, & Sacco, 1998).

Another suggestion for future research would be to investigate other methods for transforming alpha. Instead of using the transformation recommended by Bonett (2002) another possible way to transform alpha is to apply the Fisher's (1925) formula:

$$z = 1.1513 \log\left[\frac{1+r}{1-r}\right] \text{ or } z = \frac{1}{2}\ln\left[\frac{1+r}{1-r}\right]$$

It is possible that a different transformation might produce different result; however; in their Monte Carlo study using seven different approaches to average reliability, Feldt and Charter (2006) found very little difference among the averages for six of the approaches they investigated. The seventh approach they investigated was significantly different but this approach was applicable for alternative-form coefficients. They also caution the reader that these methods are for calculating the average coefficient, which should only be used as a descriptive statistic. Charter (2003) recommends using the formula:

$$r = \frac{\left[N\sum Y^2 - \left(\sum X\right)^2\right]}{\left[N\sum X^2 - \left(\sum X\right)^2\right]\left[N\sum Y^2 - \left(\sum X\right)^2\right]^{1/2}} \text{ and } \mathrm{r}_{\text{combined}} = \mathrm{r}^2$$

$$\begin{aligned} \text{where } N &= \sum n_i \\ \sum X &= \sum \left(n_i \overline{X}_i\right) \\ \sum X^2 &= \sum \left[n_i\left(\overline{X}_i^{\,2} + SD_i^2\right)\right] \\ \sum Y^2 &= \sum \left[n_i\left[\overline{X}_i^2 + \left(\sqrt{r_i}\, SD_i\right)^2\right]\right] \end{aligned}$$

and where r is the combined reliability index, $\mathrm{r}_{\text{combined}}$ is the combined reliability coefficient, n_i , $\overline{X}_i$, SD_i and r_i are the i group sample size, mean, standard deviation, and reliability coefficient, respectively. In these calculations it is assumed that the standard deviation was derived by dividing by n (for a sample standard deviation) and not dividing by n - 1 (for a population estimate). Charter (2003) points out that if the group sample size is larger than 50 the use of either type of standard deviation would be acceptable. Perhaps future research using this formula to average reliability would produce better interval estimates.

It is also possible that the use of a non-parametric sampling method such as a bootstrap method to generate the confidence intervals might provide better estimates. There are several types of bootstrap methods that might be applied to construct confidence intervals for coefficient alpha (see Hess & Kromrey, 2003). Probably the most common bootstrap method is the percentile method where samples are repeatedly drawn of size *n* with replacement from a single sample of *n* observations. Each bootstrap sample provides an estimate of coefficient alpha and the set of estimates (probably at least 1,000) would result in a distribution of point estimates of mean coefficient alpha. The 2.5 percentile and the 97.5 percentile would be the end points for a 95% confidence interval. Yuan, Guarnacci, and Hapslip (2003) investigated three methods of evaluating the distribution of the sample coefficient alpha: the existing normal-theory-based distribution, a newly proposed distribution based on fourth-order moments, and the bootstrap empirical distribution. The results of their research suggest that using the percentile method is not a good bootstrapping approach for constructing confidence intervals around an estimate of coefficient alpha. Instead, they recommend the bias corrected accelerated method that adjusts for the asymmetry in the sampling distribution and the changes in the distribution of alpha derived using the bootstrap method. In this method, the proportion of the sampling distribution that is less than the mean alpha is an estimate of asymmetry and the estimate is included in the endpoints of the 95% confidence interval. Future researchers might want to consider this method but should keep in mind that while this method may result in better interval estimates of coefficient alpha, the computation is very complex and time consuming.

Finally, this study used a fixed-effects model such that the assumption was that there is no true population variance in coefficient alpha in the RG meta-analysis. In this model variability of an infinite sample of effect size is not considered. The idea is that variance is assumed to be zero after accounting for moderators (Shadish & Haddock, 1994). This is how most RG studies have been conducted.

For a random effects model, the studies included in a meta-analysis study are really a sample from a hypothetical collection of studies such there are two sources of variance: the variability in effect size parameters and sampling error. There is a strong argument that a random effects model might be more appropriate in terms of generalizing about reliability over studies because the researcher is probably interested in generalizing the reliability of all possible studies that would use a particular measure or investigate a particular construct. Raudenbush (1994) does caution the researcher that if the number of studies used in a meta-analysis is small the random effects model would not be a good choice because the random effects variance would be a very poor estimate of the population variance.

Future research should be conducted investigating the use of random effects models to generate the interval estimates for reliability estimates in RG studies. The researcher would assume that the total variance of the observed study reliability estimates v_i^* is made up of the conditional variance v_i around the mean population reliability and the random variance $\sigma^2_{\rho_{xx}}$ such that $v_i^* = \sigma^2_{\rho_{xx}} + v_i$ (Shadish & Haddock, 1994). In this study for the construction of confidence bands, the sampling error of each estimate of score dependability index was calculated:

$$\sigma^2_{\theta k} = \frac{\mathrm{k}}{2(\mathrm{k} - 1)(\mathrm{N} - 2)}$$

where $\sigma^2_{\theta k}$ is the estimated sampling variances of z transformed r_{xx}..

The standard error used for construction of the confidence band for the mean index of score dependability was obtained as:

$$SE_{\theta} = \left(\sqrt{\sum_{k=1}^{K} \left(\frac{1}{\sigma_{\theta k}^{2}} \right)} \right)^{-1}$$

where $\sigma_{\theta k}^{2}$ is the sampling error variance for an index θ (i.e., transformed coefficient alpha) in the *k* study and the summation is across the studies included in the RG analysis.

In a random effects model the additional random variance $\sigma_{\rho_{xx}}^{2}$ would be added to the SE_{θ} and then multiplied by ±1.96 to construct the interval estimates for coefficient alpha. Because of the addition of the $\sigma_{\rho_{xx}}^{2}$ the confidence bands would be wider and therefore result in better coverage.

Thompson and Vacha-Haase (2000) suggest that RG studies have the potential to describe the stability across samples of the reliability of scores for a given scale and such an analysis also could reveal that the variation in reliability is not related to the research design factors. Before RG studies can be used to investigate these issues the design of the RG studies must first be improved to insure that the inferences made are accurate. This current study indicates that future RG researchers could use this method to describe the average reliability of scores for a given measure but should not assume that this method is appropriate for interval estimates. This research clearly indicates, contrary to the popular viewpoint, that the use of mixed models (i.e., HLM) does not necessarily alleviate the issues related to the violation of independence. More research needs to be conducted to determine the appropriate treatment of the data. This is true not only for RG studies, but for all research in general. Regardless of possible future uses and outcomes of the RG method, for these outcomes to have credibility, the RG study design must have credibility.

References

**Armsden, G. C., & Greenberg, M. T. (1987). The Inventory of Parent and Peer Attachment: Relationships to well-being in adolescence. *Journal of Youth and Adolescence, 16*, 427-454.

Bangert-Drowns, R. L. (1986). Review of developments in meta-analytic method. *Psychological Bulletin, 99,* 388-399.

Barcikowski, R. S. (1981). Statistical power with group mean as the unit of analysis. *Journal of Educational Statistics, 6,* 267-285.

*Barnes, L. L. B., Harp, D., & Jung, W. S. (2002). Reliability generalization of scores on the Spielberger State-Trait Anxiety Inventory. *Educational and Psychological Measurement*, *62*, 603-618.

**Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance. *Personnel Psychology*, *44*, 1-26.

Baugh, F. (2002). Correcting effect sizes for score reliability: A reminder that measurement and substantive issues are linked inextricably. *Educational and Psychological Measurement, 62,* 254-263.

**Beck, A. T., & Steer, R. A. (1990). *Manual for the Beck Anxiety Inventory*. San Antonio, TX: Psychological Corporation.

**Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, *4*, 53-63.

Becker, B. J. (2000). Multivariate meta-analysis. In H. E. A. Tinsley & S. Brown (Eds.), *Handbook of applied and multivariate statistics and mathematical modeling.* San Diego: Academic Press.

Becker, B. J., & Kim, J. (2002, April). *Effects of ignoring nonindependence of effect sizes on Hedge's homogeneity test.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, Louisiana.

**Bell, M., Billington, R., & Becker, B. (1985). A scale for assessment of object relations: Reliability, validity, and factorial invariance. *Journal of Clinical Psychology, 42,* 733-741.

* Beretevas, S. N., Meyers, J. L., & Leite, W. L. (2002). A reliability generalization study of the Marlowe-Crowne Social Desirability Scale. *Educational and Psychological Measurement*, *62*, 570-589.

Beretevas, S. N., & Pastor, D. A. (2003). Using mixed-effects models in Reliability Generalization studies. *Educational and Psychological Measurement*, *63*, 75-95.

**Bem, S. L. (1974). The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology*, *42,* 155-162.

Bock, R. D. (1975). *Multivariate statistical methods in behavioral research.* New York: McGrall-Hill.

Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics*, *27*, 335-340.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematics and Statistical Psychology 31,* 144-152.

**Brink, T. L., Yesavage, J. A., Lum, O., Heersema, P., Adey, M. B., & Rose, T. L. (1982). Screening tests for geriatric depression. *Clinical Gerontologist 1,* 37-44.

**Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *Minnesota Multiphasic Personality Inventory-2 (MMPI-2): Manual for administration and scoring*. Minneapolis: University of Minnesota Press.

*Capraro, R. M., & Capraro, M. M. (2002). Myers-Briggs Type Indicator score reliability across studies: A meta-analytic reliability generalization study. *Educational and Psychological Measurement*, *62*, 590-602.

*Capraro, M. M., Carpraro, R. M., & Henson, R. K. (2001). Measurement error of scores on The Mathematics Anxiety Rating Scale across studies. *Educational and Psychological Measuremen*t, *61,* 373-386.

*Caruso, J. C. (2000). Reliability Generalization of NEO personality scales. *Education and Psychological Measureme*nt, *60*, 236-254.

*Caruso, J. C., & Edwards, S. (2001). Reliability generalization of the Junior Eysenck Personality Questionnaire. *Personality and Individual Differences*, *31*, 173-184.

*Caruso, J. C., Witkiewitz, K., Belcourt-Diffloff, A., & Gottlieb, J. (2001). Reliability of scores from the Eysenck Personality Questionnaire: A Reliability Generalization (RG) study. *Educational and Psychological Measurement, 61,* 675-689.

Charter, R. A. (2003). Combining reliability coefficients: Possible application to meta-analysis and reliability generalization. *Psychological Reports, 93,* 643-647.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

**Collins, N., & Read, S. (1990). Adult attachment relationships, working models and relationship quality in dating couples. *Journal of Personality and Social Psychology, 58,* 644-683.

**Cook, C., & Heath, F. (2001). Users' perceptions of library service quality: A "LibQUAL+™" qualitative study. *Library Trends*, *49*, 548-584.

Cooper, H. (1979). Statistically combining independent studies: A meta-analysis of sex differences in conformity research. *Journal of Personality and Social Psychology, 37,* 131-146.

**Coopersmith, S. (1967). *The antecedents of self-esteem.* Palo Alto, CA: Consulting Psychologists Press.

**Costa, P. T, Jr., & McCrae, R. R. (1985). *The NEO Personality Inventory: Professional manual.* Odessa, FL: Psychological Assessment Resources.

Crocker. L., & Angina, J. (1986). *Introduction to classical and modern test theory.* Toronto: Holt Rinehart & Winston.

**Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*, 349-354.

Dawis, R. V. (1987). Scale construction. *Journal of Counseling Psychology, 34,* 481-489.

*De Ayala, R. J., Vonderharr-Carlson, D. J., & Kim, D. (2005). Assessing the reliability of the Beck Anxiety Inventory scores. *Educational and Psychological Measurement, 65,* 742-756.

*Deditius-Island, H. K., & Caruso, J. C. (2002). An examination of the reliability of scores from Zuckerman's Sensation Seeking Scales, Form V. *Educational and Psychological Measurement, 62*, 728-734.

DeShon, R. P., Ployhart, R. E., & Sacco, J. M. (1998). The estimation of reliability in longitudinal models. *International Journal of Behavioral Development, 22,* 493-515.

Dimitrov, D. M. (2002). Reliability: arguments for multiple perspectives and potential problems with generalization across studies. *Educational and Psychological Measurement, 62,* 783-801.

**Eisenberger, R., Huntington, R., Hutchison, S., & Sowa, D. (1986). Perceived organizational support. *Journal of Applied Psychology, 71,* 500-507.

**Ewing, J. A. (1984). Detecting alcoholism: The CAGE questionnaire. *Journal of the American Medical Association*, *252*, 1905-1907.

**Eysenck, H. J., & Eysenck, S. B. G. (1975). *Manual of the Eysenck Personality Questionnaire.* San Diego, CA: Educational and Industrial Testing Service.

**Eysenck, H. J., & Eysenck, S. B. G. (1994). *Manual of the Eysenck Personality Questionnaire: Comprising the EPQ-Revised (EPQ-R) and EPQ-R Short Scale*. San Diego, CA: Educational and Industrial Testing Service.

Fisher, R. A. (1925). *Statistical methods for research workers.* Edinburgh: Oliver & Boyd.

Felt, L. S., & Charter, R. A. (2006). Averaging internal consistency reliability coefficients. *Educational and Psychological Measurement*, *66*, 215-227.

Fuller, J. B., & Hester, K. (1999). Comparing the sample-weighted and unweighted meta-analysis: An applied perspective. *Journal of Management, 25,* 803-828.

**Gibson, S., & Dembo, M. (1984). Teacher efficacy: A construct validation. *Journal of Educational Psychology*, *76*, 569-582.

Glass, G. V. (1976). Primary, secondary, and meta-analysis research. *Educational Researcher, 5*(10)*,* 3-8.

Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology.* Englewood Cliffs, NJ: Prentice Hall.

Glass, G. V., Peckman, P., & Sanders, J. (1972). Consequences of failure to meet assumptions underlying for the fixed effects analyses of variance and covariance. *Review of Educational Research, 42,* 237-288.

Glass, G. V., & Smith, M. L. (1979*).* Meta-analysis of research on class size and achievement. *Educational Evaluation and Policy Analysis*, *1,* 2-16.

Gleser, L. J., & Olkin, I. (1994). Stochastically dependent effect sizes. In H. Cooper & L. Hedges (Eds.), *The handbook of research synthesis* (pp. 339-355). New York: Russell Sage.

Greenhouse, J. B., & Iyengar, S.(1994). Sensitivity analysis and diagnostics. In H. Cooper & L. Hedges (Eds.), *The handbook of research synthesis* (pp. 383-398). New York: Russell Sage.

Gronlund, N. E., & Linn, R. L. (1990). *Measurement and evaluation in teaching.* New York: Macmillan.

**Guskey, T. R. (1981, April). *Differences in teachers' perceptions of the causes of positive versus negative student achievement outcomes*. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles. (ERIC Document Reproduction Service No. ED 200 624)

Hanson, B. A., & Beguin, A. A. (1999, April). *Obtaining a common scale for IRT item parameters using separate versus concurrent estimation in the common item nonequivalent groups equating method.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

*Hanson, W. E., Curry, K. T., & Bandalos, D. L. (2002). Reliability generalization of Working Alliance Inventory scale scores. *Educational and Psychological Measurement*, *62*, 659-673.

Hedges, L. V. (1982). Statistical methodology in meta-analysis. *ERIC Clearinghouse on Test, Measures, and Evaluation, Educational Testing Services*, Princeton, NJ

Hedges, L. V., & Olkin, I. (1985). *Statistical Methods for Meta-Analysis.* San Diego, CA: Academic Press, Inc.

*Helms, J. E. (1999). Another meta-analysis if the White Racial Identity Attitude Scale's Cronbach alphas: Implication for validity. *Measurement and Evaluation in Counseling and Development*, 32, 122-137.

**Helms, J. E., & Carter, R. T. (1990). Development of the White Racial Identity Attitude Inventory. In J. E. Helms (Ed.), *Black and White racial identity: Theory, research and practice* (pp. 67-80). Westport, CT: Greenwood Press.

*Hellman, C. M., Fuqua, D. R., & Worley, J. (2006). A reliability generalization study on the Survey of Perceived Organizational Support. *Educational and Psychological Measurement*, *66*, 631-642.

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

*Henson, R. K., Kogan, L. R., & Vacha-Haase, T. (2001). A reliability generalization study of the Teacher Efficacy Scales and related instruments. *Educational and Psychological Measurement, 61*, 404-420.

*Henson, R. K., & Hwang, D. (2002). Variability and prediction of measurement error in Kolb's Learning Style Inventory scores: A reliability generalization study. *Educational and Psychological Measurement*, *62*, 712-727.

*Henson, R. K., & Thompson, B. (2001, April). *Characterizing measurement error in test scores across studies: A tutorial on conducting "Reliability Generalization" Analysis.* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA

Hess, M. R., & Kromrey, J. D. (2003, April). *Confidence intervals for standardized mean differences: An empirical comparison of bootstrap methods under non-normality and heterogeneous variances.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, Il.

Hewitt-Gervais, C. M., & Kromrey, J. D. (1999, April). *Analysis strategies for conducting F-tests with non-independent observations: An empirical investigation.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Hopkins, K. D. (1998). *Educational and psychological measurement and evaluation.* Boston: Allyn & Bacon.

**Horvath A. O., & Greenberg L. S. (1989). Development and validation of the Working Alliance Inventory. J*ournal of Counseling Psychology, 36*, 223–233

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

**Izard, C. E., Libero, D. Z., Putnam, P. & Haynes, O. M. (1993). Stability of emotion experiences and their relations to traits of personality. *Journal of Personality and Social Psychology, 64,* 847-860.

Kenny, D. A., & Judd, C. M. (1986). Consequences of violating the independence assumption in analysis of variance. *Psychological Bulletin, 99,* 422-431.

**Kenny, M. E. (1987). The extent and function of the parental attachment among first-year college students. *Journal of Youth and Adolescence, 16,* 17-27.

*Kieffer, K. M., & Reese, R. J. (2003). A reliability generalization study of the Geriatric Depression Scale. *Educational and Psychological Measurement*, *63,* 279-295.

**Kolb, D. A. (1976). *Learning Style Inventory technical manual*. Boston: McBer.

Kreft, I., & de Leeuw, J., (1998). *Introducing multilevel modeling.* London: Sage

Kromrey, J. D., & Dickinson, W. B. (1996). Detecting unit of analysis problems in nested methods: Statistical power and Type I error rates of the F test for groups-within-treatments effects. *Educational and Psychological Measurement, 56,* 215-231.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2,* 151-160.

Landman, J. T., & Dawes, R. M. (1982). Psychotherapy outcomes: Smith and Glass' conclusions stand up to scrutiny. *American Psychologist, 37,* 504-516.

*Lane, G. G., White, A. E., & Henson, R. K. (2002). Expanding reliability generalization methods with KR-21 estimates: An RG study of the Coopersmith Self-Esteem Inventory. *Educational and Psychological Measurement*, *62*, 685-711.

*Leach, L. F., Henson, R. K., Odom, L. R., & Cagle, L. S. (2006). A reliability generalization study of the Self-Description s: An RG study of the Coopersmith Self-Esteem Inventory. *Educational and Psychological Measurement*, *62*, 685-711.

Mansfield, R. S., & Busse, R. R. (1977). Meta-analysis of research: A rejoinder to Glass. *Educational Researcher, 6*(9)*,* 3-4.

**Midgley, C., Maehr, M., Hicks, L., Roeser, R., Urdan, T., & Anderman, E. (1997). *Patterns of Adaptive Learning Survey (PALS)*. Ann Arbor: University of Michigan.

**Myers, I. B., & McCaulley, M. H. (1985). *Manual: A guide to the development and use of the Myers-Briggs Type Indicator*. Palo Alto, CA: Consulting Psychologists Press.

**Neugarten, B. L., Havighurst, R. J. & Tobin, S. S. (1961). The measurement of life satisfaction. *Journal of Gerontology, 16,* 134-143.

*Nilsson, J. E., Schmidt, C. K., & Meek, W. D. (2002). Reliability generalization: An examination of the Career Decision-Making Self-Efficacy Scale. *Educational and Psychological Measurement*, *62*, 647-658.

*O'Rourke, N. (2004). Reliability generalization of responses by care providers to the Center for Epidemiologic Studies—Depression Scale. *Educational and Psychological Measurement*, *64*, 973-990.

**Parker, G., Tulping, H., & Brown, L. B. (1979). A parental bonding instrument. *British Journal of Medical Psychology 52*, 1–10.

Pedhazur, E. (1982). *Multiple regression in behavioral research.* New York: Holt, Rinehart and Winston.

Pedhazur, E., & Schmelkin, L. (1991). *Measurement, method, and analysis: An integrated approach.* Hillsdale, NJ: Erlbaum.

**Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1,* 385-401.

Raudenbush, S. W. (1994). Random effects models. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301-321). New York, NY: Russell Sage Foundation.

Raudenbush, S. W., Becker, B. J., & Kalaian, H. (1988). Modeling multivariate effect sizes. *Psychological Bulletin, 103*, 111-120

Raudenbush, S. W., & Bryk, A. S. (1987). Examining correlate of diversity. *Journal of Educational Statistics, 12,* 241-269.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and data analysis methods* (2nd Edition). Thousand Oaks, CA: Sage.

*Reese, R. J., Kieffer, K. M., & Briggs, B. K. (2002). A reliability generalization study of select measures of adult attachment style. *Educational and Psychological Measurement*, *62*, 619-646.

**Reynolds, C. R., & Paget, K. D. (1983). National normative and reliability data for the Revised Children’s Manifest Anxiety Scale. *School Psychology Review, 12,* 324-336.

**Reynolds, C. R., & Richmond B. O. (1985). *Revised Children’s Manifest Anxiety Scale (RCMAS) manual*. Los Angeles: Western Psychological Services.

**Richardson, F. C., & Suinn, R. M. (1972). The Mathematics Anxiety Rating Scale: Psychometric data. *Journal of Counseling Psychology, 19*, 551-554.

**Riggs, I., & Enochs, L. (1990). Toward the development of an elementary teacher’s science teaching efficacy belief instrument. *Science Education*, *74*, 625-638

Robey, R., & Barcikowski, R. (1992) Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematics and Statistical Psychology, 45,* 283-288.

Romano, J. L., & Kromrey, J. D. (2002, April). *The "RG sausage's" missing ingredients:investigating the validity of the Reliability Generalization study method.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, Louisiana.

Romano, J. L., & Kromrey, J. D. (2004, April). *"Spicing up" the Reliability Generalization study method: investigating the inferences using internal consistency estimates of reliability.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

**Rose, J. S., & Medway, F. J. (1981). Measurement of teachers' beliefs in their control over student outcome. *Journal of Educational Research*, *74*, 185-190.

Rosenthal, R. (1979). The "file-drawer problem" and tolerance for null results. *Psychological Bulletin, 86,* 638-641.

Rosenthal, R. (1994). Parametric measures of effect sizes. In H. Cooper., & L.V. Hedges (Eds.), *The Handbook of Research Synthesis* (pp. 231-260). New York: Russell Sage Foundation.

Rosenthal, R., & Rubin, D. B. (1986). Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological Bulletin, 99,* 400-406.

*Ross, M. E., Blackburn, M., & Forbes, S. (2005). Reliability Generalization of the Patterns of the Adaptive Learning Survey goal orientation scales. *Educational and Psychological Measurement, 65,* 451-464.

*Ryngala, D. J., Shields, A. L., & Caruso, J. C. (2005). Reliability Generalization of the Revised Children's Manifest Scale. *Educational and Psychological Measurement, 65,* 259-271.

**Saunders, J. B., Aasland, O. G., Babor, T. F., de la Fuente, J. R., & Grant, M. (1993). Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption. II. *Addiction*, *88*, 791-804.

Sawilosky, S. S. (2000). Psychometrics versus datametrics: Comment on Vacha-Haase's ”reliability generalization” method and some ERM editorial policies. *Educational and Psychological Measurement, 60,* 157-173.

Scariano, S. M., & Davenport, J. M. (1987). The effects of violations of independence assumptions in a one-way ANOVA. *The American Statistician, 41,* 123-129.

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.

Shadish, W. R., & Haddock, C. K. (1994). Combing estimates of effect size. In H. Cooper & L. V. Hedges (Eds.), *The Handbook of Research Synthesis* (pp. 261-281). New York, NY: Russell Sage Foundation.

*Shields, L., & Caruso, J. C. (2003). Reliability generalization of the Alcohol Use Disorders Identification Test. *Educational and Psychological Measurement, 63,* 404-413.

*Shields, L., & Caruso, J. C. (2004). A reliability induction and reliability generalization study of the Cage Questionnaire. *Educational and Psychological Measurement, 64,* 254-270.

Smith, M. L., Glass, G. V., & Miller, T. I. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32,* 752-760.

Smith, M. L., Glass, G. V., & Miller, T. I. (1980). *The benefits of psychotherapy,* Baltimore, MD: Johns Hopkins University Press.

**Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. (1970). *Manual for the State-Trait Anxiety Inventory.* Palo Alto, CA: Consulting Psychologists Press.

Stevens, J. P. (1999). *Intermediate statistics: A modern approach* (2nd ed.). Mahwah, NJ: Erlbaum.

**Taylor, K. M., & Betz, N. E. (1983). Applications of self-efficacy theory to the understanding and treatment of career indecision. *Journal of Vocational Behavior, 37*, 17-31.

*Thompson, B., & Cook, C. (2002). Stability of the reliability of LibQUAL+$^{\text{TM}}$ scores: A reliability generalization meta-analysis study. *Educational and Psychological Measurement, 62,* 735-743.

Thompson, B., & Snyder, P. A. (1998). Statistical significance and reliability analyses in recent *JCD* research articles. *Journal of Counseling and Development, 76,* 436-441.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics *is* datametrics: The test is not reliable. *Educational and Psychological Measurement, 60,* 174-195.

Tracz, S. M., Elmore, P. B., & Pohlmann, J. T. (1992). Correlational meta-analysis: independent and nonindependent cases. *Educational and Psychological Measurement*, *52*, 879-888.

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

*Vacha-Haase, T., Kogan, L., Tani, C. R., & Woodall, R. A. (2001). Reliability generalization: Exploring reliability coefficients of MMPI clinical scales scores. *Educational and Psychological Measurement*, *61*, 45-59.

Vacha-Haase, T., Kogan, L. R., & Thompson. B. (2000). Sample composition and variabilities in published studies versus the in test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement, 60,* 509-522.

Vacha-Haase, T., Ness, C., Nilsson, J., & Rettz, D. (1999). Practices regarding reporting of reliability coefficients: A review if three journals. *Journal of Experimental Education, 67,* 335-341.

*Vacha-Haase, T., Tani, C. R., Kogan, L. R., Woodall, R. A. & Thompson, B. (2001). Reliability generalization: Exploring reliability variations on MMPI/MMPI-2 Validity scale scores. *Assessment, 8,* 391-401.

*Viswesvaran, C., & Ones, D. (2000). Measurement error in "Big Five Factors" personality assessment: Reliability Generalization across studies and measures. *Educational and Psychological Measurement*, *60*, 24-235.

*Wallace, K. A., & Wheeler, A. J. (2002) Reliability generalization of the Life Satisfaction Index. *Educational and Psychological Measurement*, *62*, 674-684.

Wilkinson, L., & American Psychological Association (APA) Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594-604.

Wortman, P. M., & Bryant, F. B. (1984). Methodological issues in the meta-analysis of quasi-experiments. *New Directions for Program Evaluation, 24,* 5-24.

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

*Yin, P., & Fan, X. (2000). Assessing the reliability of Beck Depression Inventory scores: Reliability Generalization across studies. *Educational and Psychological Measurement*, *61*, 201-223.

**Yesavage, J. A., Brink, T. L, Rose, T. L, Lum, O., Huang, V., Adey, M. B., & Leirer, V. O. (1983). Development and validation of a geriatric depression screening scale: A preliminary report. *Journal of Psychiatric Research, 17,* 37-49.

*Youngstrom, E. A., & Green, K. W. (2003). Reliability generalization of self-report of emotions when using the Differential Emotions Scale. *Educational and Psychological Measurement*, *63*, 279-295.

**Zukerman, M. (1994). *Behavior expressions and biosocial bases of sensation seeking.* New York: Cambridge University Press.

Note: * indicates RG study and ** indicates scale evaluated in an RG study

Appendices:

Appendix A: SAS Code for Monte Carlo Simulation

```
options ls=132;
proc printto log= 'C:\rg\mylogyes';
proc printto print='C:\rg\IC01r90n10k15.txt';
* +----------------------------------------------------------------------------------------------+
   RG_Block Alpha.SAS: Simulates conditions for an entire block of the design matrix

   30 September 2001: Modified the weights used in weighted means for rxx and SEM.
                      The sample value of the statistic is no longer a part of the weight.
   10 July 2003: Added subroutine for analysis of coefficient alpha
   22 July 2003: Simplified the output section: matrices instead of scalars
                         Simplified subroutines for weighted and unweighted mean calculations
  +----------------------------------------------------------------------------------------------+;

data iosif;
 input item_no a_3pl b_3pl c_3pl;
 poolid = _n_;

*if item_no < 4; *3 items for .3;
*if item_no<7;  *6 items for .5;
*if item_no < 12; *11 items for .7;
*if item_no <51; *50 items for .90;
cards;
1     0.642 -2.522      0.187
2     0.806 -1.902      0.149
3     0.956 -1.351      0.108
4     0.972 -1.092      0.142
5     1.045 -0.234      0.373
6     0.834 -0.317      0.135
7     0.614 0.037 0.172
8     0.796 0.268 0.101
9     1.171 -0.571      0.192
10    1.514 0.317 0.312
11    0.842 0.295 0.211
12    1.754 0.778 0.123
13    0.839 1.514 0.17
14    0.998 1.744 0.057
15    0.727 1.951 0.194
16    0.892 -1.152      0.238
```

```
17    0.789 -0.526      0.115
18    1.604 1.104 0.475
19    0.722 0.961 0.151
20    1.549 1.314 0.197
21    0.7       -2.198  0.184
22    0.799 -1.621      0.141
23    1.022 -0.761      0.439
24    0.86  -1.179      0.131
25    1.248 -0.61 0.145
26    0.896 -0.291      0.082
27    0.679 0.067 0.161
28    0.996 0.706 0.21
29    0.42  -2.713      0.171
30    0.977 0.213 0.28
31    1.257 0.116 0.209
32    0.984 0.273 0.121
33    1.174 0.84  0.091
34    1.601 0.745 0.043
35    1.876 1.485 0.177
36    0.62  -1.208      0.191
37    0.994 0.189 0.242
38    1.246 0.345 0.187
39    1.175 0.962 0.1
40    1.715 1.592 0.096
41    0.769 -1.944      0.161
42    0.934 -1.348      0.174
43    0.496 -1.348      0.328
44    0.888 -0.859      0.199
45    0.953 -0.19 0.212
46    1.022 -0.116      0.158
47    1.012 0.421 0.288
48    1.605 1.377 0.12
49    1.009 -1.126      0.133
50    1.31  -0.067      0.141
51    0.957 0.192 0.194
52    1.269 0.683 0.15
53    1.664 1.017 0.162
54    1.511 1.393 0.123
```

```
55    0.561 -1.865      0.24
56    0.728 -0.678      0.244
57    1.665 -0.036      0.109
58    1.401 0.117 0.057
59    1.391 0.031 0.181
60    1.259 0.259 0.229
61    0.804 -2.283      0.192
62    0.734 -1.475      0.233
63    1.523 -0.995      0.175
64    0.72  -1.068      0.128
65    0.892 -0.334      0.211
66    1.217 -0.29 0.138
67    0.891 0.157 0.162
68    0.972 0.256 0.126
69    1.206 -0.463      0.269
70    1.354 0.122 0.211
71    0.935 -0.061      0.086
72    1.438 0.692 0.209
73    1.613 0.686 0.096
74    1.199 1.097 0.032
75    0.786 -1.132      0.226
76    1.041 0.131 0.15
77    1.285 0.17  0.077
78    1.219 0.605 0.128
79    1.473 1.668 0.187
80    1.334 0.53  0.075
81    0.965 -1.862      0.152
82    0.71  -1.589      0.138
83    0.523 -1.754      0.149
84    1.134 -0.604      0.181
85    0.709 -0.68 0.064
86    0.496 -0.443      0.142
87    0.979 0.181 0.124
88    0.97  0.351 0.151
89    0.524 -2.265      0.22
90    0.944 -0.084      0.432
91    0.833 0.137 0.202
92    1.127 0.478 0.199
```

```
93    0.893 0.496 0.1
94    1.215 0.867 0.076
95    1.079 -0.486      0.264
96    0.932 0.45  0.259
97    1.141 0.344 0.071
98    1.068 0.893 0.153
99    1.217 1.487 0.069
100 1.310   1.186 0.153
;
proc iml symsize = 500;
* +---------------------------------------------------+
    Define parameters for execution of the simulation
  +---------------------------------------------------+;
   replicat=10000;  * N of meta-analyses to simulate This value will be set to 10,000;
   icc=.01;
  *N1 njs = 10;     * average sample size in study;
  *N2 njs = 50;
  *N3 njs = 100;
  *N4 njs = 250;
  *N5 njs = 500;
  *N6;* njs = 1500;

    mu1=0;    * Pop mean;
sds = 1;

*+----------------------------------------------------------------------+
3, May 2005 Subroutine to calculate a mean rxx where ind is violated

* +-------------------------------------------------------------------------+
    Subroutine to calculate vector of variabilities for coefficient alpha,

     Both original alpha metric and Fishers z are used
   Inputs to the subroutine are
     ri_by_k - a matrix  of sample alpha estimates where
     ri is the number of rows(i.e. #alphas per study)
     k is the nunmber of columns (i.e. # of studies)
     items - number of items on the test (scalar value)
     N_vec_mtx  - matrix  of sample sizes corresponding to each reliability
   Outputs are (29, April 2005 some of these variables are not needed for J9 dis)
            Z_w_mean -  weighted mean Fisher Z
            SE_Z = Standard error of mean Fisher Z
  +------------------------------------------------------------------------+;
start calc_alphaVI(ri_by_k,items,N_vec_mtx,Z_W_mean,SE_Z);
* note J9 deleted values from sub routine that are not used in new study;

* calculate variance for each reliability estimate, Fisher Z and variance of the Z;
   * Have to chance this so that I indexing using rows and columns of a matrix
that is nr by n_studies;
*print 'IV';
*print ri_by_k;
*print n_vec;
k=ncol(ri_by_k); * ie number of studies;
q = nrow(ri_by_k); * ie number of reliabilties per study;

  Z_alpha=J(q,k,0);
  var_Z =J(q,k,0);

  do i = 1 to k;
   do   v = 1 to q;
      * +-----------------------------------------------+
         Be sure the Rxx values are between .01 and .99
      * +-----------------------------------------------+;
            if ri_by_k[v,i] > .99 then ri_by_k [v,i] = .99;
            if ri_by_k[v,i] < -.99 then ri_by_k[v,i] = -.99;

      * +----------------------------------------------------+
         Fisher Z transformation, from Bonett, 2002
      * +----------------------------------------------------+;
            Z_alpha[v,i] = log(1-abs(ri_by_k[v,i]));
            if ri_by_k[v,i] <0 then Z_alpha[v,i] = Z_alpha[v,i]* -1;
         *new code added for N_vec_matrix;
  * N_vec[1,i] = n_vec[1,i] + N_vec_mtx[v,i];
      var_Z[v,i] = (2#items)/((items - 1) # (N_vec_mtx[v,i] - 2));

end; * q end;
```

```
 end; * k end;
      * +----------------------------------------------------+
         Calculate weighted mean alpha and mean Z
      * +----------------------------------------------------+;
*Rxx_w_mean = 0;
      Z_W_mean= 0;
 *    Sum_wt = 0;
Sum_wtz = 0;

      do i = 1 to k;
       do v = 1 to q;
  *          Rxx_w_mean = Rxx_w_mean + ri_by_k[i,1]/var_alpha[i,1];
   *         Sum_wt = sum_wt + var_alpha[i,1]##-1;
      Z_W_mean = Z_W_mean + Z_alpha[v,i]/var_Z[v,i];
      Sum_wtz = sum_wtz + var_Z[v,i]##-1;
        end; * q end;
 end; * k end;
 *print z_w_mean sum_wtz;
      *Rxx_w_mean = Rxx_w_mean/sum_wt;
      Z_W_mean = Z_W_mean/sum_wtz;
      *print z_w_mean;
      * +-------------------------------------------------------+
         Calculate standard errors of the mean alpha and mean Z
      * +-------------------------------------------------------+;
      *SE_Rxx = sqrt(sum_wt##-1);
      SE_Z = sqrt(sum_wtz##-1);
      * +----------------------------------------------------+
         Calculate unweighted mean alpha and mean Z
      * +----------------------------------------------------+;
 *    Rxx_U_mean = (J(1,k,1)*ri_by_k)/k;
 *    Z_U_mean = (J(1,k,1)*Z_alpha)/k;
finish;
*+-----------------------------------------------------------------
            End of Vio Ind subroutine
            Begining subroutine for calc mean for each study
*+-----------------------------------------------------------------;
*@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@;
*+----------------------------------------------------------------------+
                3, May 2005 Subroutine to calculate a mean rxx
                      where the mean of each study is
                             the unit of analysis
* +--------------------------------------------------------------------------+
    Subroutine to calculate vector of variabilities for coefficient alpha,

     Both original alpha metric and Fishers z are used
   Inputs to the subroutine are
ri_by_k - a matrix  of sample alpha estimates where
     ri is the number of rows(i.e. #alphas per study)
     k is the nunmber of columns (i.e. # of studies)
     items - number of items on the test (scalar value)
     n_vec  - vector of sample sizes corresponding to each reliability

   Outputs are (29, April 2005 some of these variables are not needed for J9 dis)
            Z_w_mean -  weighted mean Fisher Z
            SE_Z = Standard error of mean Fisher Z
  +-------------------------------------------------------------------------+;
start calc_kalpha_mean(ri_by_k,items,N_vec_mtx,Z_W_mean,SE_Z);
* note J9 deleted values from sub routine that are not used in new study;

* calculate variance for each reliability estimate, Fisher Z and variance of the Z;
   * Have to chance this so that I indexing using rows and columns of a matrix
that is nr by n_studies q by k);
*print 'In CALC_KALPHA_MEAN';

k=ncol(ri_by_k);
q = nrow(ri_by_k);
   Z_alpha=J(q,k,0);
  var_Z =J(q,k,0);
mean_vec= J(1,k,0);
n_vec=J(1,k,0);
  do i = 1 to k;
   do  v = 1 to q;

      * +-----------------------------------------------+
         Be sure the Rxx values are between .01 and .99
      * +-----------------------------------------------+;
```

```
            if ri_by_k[v,i] > .99 then ri_by_k [v,i] = .99;
            if ri_by_k[v,i] < -.99 then ri_by_k[v,i] = -.99;
      * +---------------------------------------------------+
         Fisher Z transformation, from Bonett, 2002
      * +---------------------------------------------------+;
            Z_alpha[v,i] = log(1-abs(ri_by_k[v,i]));
            if ri_by_k[v,i] <0 then Z_alpha[v,i] = Z_alpha[v,i]* -1;
      mean_vec[1,i] = mean_vec[1,i]+ Z_alpha[v,i];
            N_vec[1,i] = n_vec[1,i] + N_vec_mtx[v,i];
end; * q end;
 end; * k end;

   mean_vec= mean_vec/q;

N_vec= n_vec/q;
   do i= 1 to k;
            var_Z[1,i] = (2#items)/((items - 1) # (N_vec[1,i] - 2));
    end; *k end;

      * +---------------------------------------------------+
         Calculate weighted mean alpha and mean Z
      * +---------------------------------------------------+;
      Z_W_mean= 0;
      Sum_wtz = 0;

      do i = 1 to k;

      Z_W_mean = Z_W_mean + mean_vec[1,i]/var_Z[1,i];
      Sum_wtz = sum_wtz + var_Z[1,i]##-1;
      end; * k end;

      Z_W_mean = Z_W_mean/sum_wtz;
      * +------------------------------------------------------+
         Calculate standard errors of the mean alpha and mean Z
      * +------------------------------------------------------+;
      SE_Z = sqrt(sum_wtz##-1);
*print z_w_mean se_z;
finish;
```

```
*+-------------------------------------------------------------------+
                              End of mean subroutine
                        Beginning of Median subroutine
where the median of each study is the unit of analysiss
*+---------------------------------------------------------------------+;
*@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@;
* 4, June 2005 Subroutine to calculate a median rxx

* +--------------------------------------------------------------------------+
   Subroutine to calculate vector of variabilities for coefficient alpha
   Both original alpha metric and Fishers z are used
   *Inputs to the subroutine are
     ri_by_k - a matrix  of sample alpha estimates where
     ri or q is the number of rows(i.e. #alphas per study)
     k is the nunmber of columns (i.e. # of studies)
     items - number of items on the test (scalar value)
    N_vec_mtx  - matrix  of sample sizes corresponding to each reliability

   Outputs are (29, April 2005 some of these variables are not needed for J9 dis)
            Z_w_mean -  weighted mean Fisher Z
            SE_Z = Standard error of mean Fisher Z
  +-------------------------------------------------------------------------+;
start calc_kalpha_Med(ri_by_k,items,N_vec_mtx,Z_W_mean,SE_Z);
* note J9 deleted values from sub routine that are not used in new study;

* calculate variance for each reliability estimate, Fisher Z and variance of the Z;
   * Have to chance this so that I indexing using rows and columns of a matrix
that is nr by n_studies;
*print 'In calc_kalpha_Md';
k=ncol(ri_by_k);
q = nrow(ri_by_k);
  var_alpha=J(q,k,0);
  Z_alpha=J(q,k,0);
  var_Z =J(q,k,0);
Md_vec= J(1,k,0);
 N_vec = J(1,k,0) ;
  do i = 1 to k;
```

```
   do  v = 1 to q;
      * +----------------------------------------------+
         Be sure the Rxx values are between .01 and .99
      * +----------------------------------------------+;
            if ri_by_k[v,i] > .99 then ri_by_k [v,i] = .99;
            if ri_by_k[v,i] < -.99 then ri_by_k[v,i] = -.99;
      * +---------------------------------------------------+
         Fisher Z transformation, from Bonett, 2002
      * +---------------------------------------------------+;
            Z_alpha[v,i] = log(1-abs(ri_by_k[v,i]));
            if ri_by_k[v,i] <0 then Z_alpha[v,i] = Z_alpha[v,i]* -1;

   end; * q end;
 end; * k end;
      * +---------------------------------------------------+
         Compute upper and lower endpoints of the confidence
       interval suggested by Feldt et al.. (1987)
      * +---------------------------------------------------+;
if q = 1 | q = 3 then do;
      w = (q+1)/2;
      *print w;
      do i = 1 to k;
            r= rank(Z_alpha[,i]);
            *print r w;
            do  v = 1 to q;
                  if r[v]=w then Md_vec[1,i] = Z_alpha[v,i];
            end;
      end;
end;

if q = 2 | q = 10 | q = 50 then do;*FIXED THIS;
      m1 = q/2;
      m2= (q+2)/2;
      do i = 1 to k;
            r= rank(Z_alpha[,i]);
            *print r m1 m2;
* BEGIN NEW PART OF CODE;
            do  v = 1 to q;
```

```
                  if r[v]=m1 then Md_part1 = Z_alpha[v,i];
                  if r[v]=m2 then Md_part2 = Z_alpha[v,i];
            end;
            *print Md_part1 Md_part2;
            Md_vec[1,i] = (Md_part1 + Md_part2)/2;
* END NEW PART OF CODE;
      end;
end; *ADDED THIS END;
*print 'vector of medians' Md_vec;
do i=1 to k;                               *+-------code for mean of sample size-------+;
do v = 1 to q;
   N_vec[1,i] = n_vec[1,i] + N_vec_mtx[v,i];
   end;
end;
N_vec= n_vec/q;
   do i= 1 to k;
            var_Z[1,i] = (2#items)/((items - 1) # (N_vec[1,i] - 2));
 end; * k end;
      * +---------------------------------------------------+
         Calculate weighted mean alpha and mean Z
      * +---------------------------------------------------+;
      Z_W_mean= 0;
      Sum_wtz = 0;

      do i = 1 to k;

      Z_W_mean = Z_W_mean + Md_vec[1,i]/var_Z[1,i];
      Sum_wtz = sum_wtz + var_Z[1,i]##-1;
      end; * k end;

      Z_W_mean = Z_W_mean/sum_wtz;
      SE_Z = sqrt(sum_wtz##-1);
      *print z_w_mean se_z;
      finish;
*+---------------------------------------------------------------------------+
                              End Median Rutine begin Random routine
+----------------------------------------------------------------------------+;
 *@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@;
```

```
* 4, June 2005 Subroutine to calculate a random rxx from each study
      where the unit of analysis is a randomly selected rxx from each of the k
      studies.
* +-------------------------------------------------------------------------+
   Subroutine to calculate vector of variabilities for coefficient alpha

    Both original alpha metric and Fishers z are used
   *Inputs to the subroutine are
     ri_by_k - a matrix  of sample alpha estimates where
     ri or q is the number of rows(i.e. #alphas per study)
     k is the nunmber of columns (i.e. # of studies)
     items - number of items on the test (scalar value)
     N_vec_mtx  - matrix of sample sizes corresponding to each reliability

   Outputs are (29, April 2005 some of these variables are not needed for J9 dis)
            Z_w_mean -  weighted mean Fisher Z
            SE_Z = Standard error of mean Fisher Z
  +------------------------------------------------------------------------+;
start calc_kalpha_rand(ri_by_k,items,N_vec_mtx,Z_W_mean,SE_Z);

* calculate variance for each reliability estimate, Fisher Z and variance of the Z;
   * Have to chance this so that I indexing using rows and columns of a matrix
that is nr by n_studies;

k=ncol(ri_by_k);
q = nrow(ri_by_k);
  var_alpha=J(q,k,0);
  Z_alpha=J(q,k,0);
  var_Z =J(q,k,0);
   N_vec = J(1,k,0) ;
rand_vec= J(1,k,0);
  do i = 1 to k;
   do  v = 1 to q;
      * +-----------------------------------------------+
         Be sure the Rxx values are between .01 and .99
      * +-----------------------------------------------+;
            if ri_by_k[v,i] > .99 then ri_by_k [v,i] = .99;
            if ri_by_k[v,i] < -.99 then ri_by_k[v,i] = -.99;
   end; * q end;
 end; * k end;

      * +---------------------------------------------------+
         Compute upper and lower endpoints of the confidence
       interval suggested by Feldt et al.. (1987)
      * +---------------------------------------------------+;

s=J(q,1,0);

do i= 1 to k;
      seed1 = round(1000000*ranuni(0));
      do v = 1 to q;
            s[v,1]= rannor(seed1);
      end;
      r=rank(s);
      *print s r;
do v = 1 to q;
            if r[v] = q then do;
            rand_vec[1,i] = ri_by_k[v,i];
            N_vec[1,i]=n_vec_mtx[v,i];
            end;
      end; * q end;
 end; * k end;


   do i= 1 to k;
   *do v = 1 to q;
*     upper = 1 - (1 - mean_vec[1,i])#FINV(.05,(n_vec[v,1] - 1),((n_vec[v,1] - 1) # (items - 1)));
 *    lower = 1 - (1 - mean_vec[1,i])#FINV(.95,(n_vec[v,1] - 1),((n_vec[v,1] - 1) # (items - 1)));

            * +--------------------------------------------------------+
         Use the width of the confidence interval to compute an
         equivalent variance for alpha. If alpha was normally
         distributed, this would be the SE.
      * +--------------------------------------------------------+;

      *var_alpha[v,i]= (abs(upper - lower)/(2#1.96))##2;
      * +----------------------------------------------------+
         Fisher Z transformation, from Bonett, 2002
      * +----------------------------------------------------+;
            Z_alpha[1,i] = log(1-abs(rand_vec[1,i]));
            if rand_vec[1,i] <0 then Z_alpha[1,i] = Z_alpha[1,i]* -1;
      *      N_vec[1,i] = n_vec[1,i] + N_vec_mtx[v,i];   *new code added for N_vec;
            var_Z[1,i] = (2#items)/((items - 1) # (N_vec[1,i] - 2));
*print Z_alpha;
*end; * q end;
 end; * k end;
      * +----------------------------------------------------+
         Calculate weighted mean alpha and mean Z
      * +----------------------------------------------------+;
*Rxx_w_mean = 0;
      Z_W_mean= 0;
 *    Sum_wt = 0;
      Sum_wtz = 0;
      do i = 1 to k;


  *         Rxx_w_mean = Rxx_w_mean + ri_by_k[i,1]/var_alpha[i,1];
   *        Sum_wt = sum_wt + var_alpha[i,1]##-1;
      Z_W_mean = Z_W_mean + Z_alpha[1,i]/var_Z[1,i];
      Sum_wtz = sum_wtz + var_Z[1,i]##-1;
      end; * k end;

      *Rxx_w_mean = Rxx_w_mean/sum_wt;
      Z_W_mean = Z_W_mean/sum_wtz;
      * +-------------------------------------------------------+
         Calculate standard errors of the mean alpha and mean Z
      * +-------------------------------------------------------+;
      *SE_Rxx = sqrt(sum_wt##-1);
      SE_Z = sqrt(sum_wtz##-1);
      finish;

START IRT3PL (THETA, A, B, C, SEED1, SUM_P, SCORE3PL);
 * +---------------------------------------------------------------+
    Subroutine to compute probabilities of correct responses under
    3PL model.

    INPUTS:  Theta = scalar ability
             A, B, C = column vectors of item parameters
             SEED1 = seed for random number generator

    OUTPUTS: SUM_P = true number correct score (sum of true p-values)
             SCORE3PL = row vector of scored items (0,1)
   +---------------------------------------------------------------+;
      n_items = NROW(A);
      SUM_P = 0;
      SCORE3PL = J(1,n_items,0);
      do i = 1 to n_items;
        AVAL = -1.702 * A[i,1];
        DAB = AVAL * (THETA - B[i,1]);
        IF DAB > 120 THEN P = C[i,1];
        IF DAB < -100 THEN P = .99999;
        IF DAB >= -100 & DAB<=120 then do;
           DIV = 1 + EXP(DAB);
           P = C[i,1] + (1.0 - C[i,1])/DIV;
        END; * end DAB;
        RANVAR = RANUNI(SEED1);
        IF RANVAR <= P THEN SCORE3PL[1,i] = 1;
        IF RANVAR > P THEN SCORE3PL[1,i] = 0;
            SUM_P = SUM_P + P;
      end;  *end n_items;
FINISH;

* +--------------------------------------------+
   Main program
   Generates samples, calls subroutines,
   computes means and confidence band coverage.
  +--------------------------------------------+;

* +-------------------------------------+
   Reading in the item pool information
  +-------------------------------------+;
  use iosif;
  read all var {a_3pl} into ta_3PL;
  read all var {b_3pl} into tb_3PL;
  read all var {c_3pl} into tc_3PL;
  read all var {poolid} into poolid;
*print ta_3PL tb_3pl tc_3PL;


*do pop_sds = 1 to 4;
*     if pop_sds = 1 then sds = 1;
*     if pop_sds = 2 then sds = 2;
*     if pop_sds = 3 then sds = 4;
*     if pop_sds = 4 then sds = 8;
do rel_items =1 to 1;
 * if rel_items =1 then true_alpha = .33;
 *if rel_items = 1  then true_alpha = .54;
      *if rel_items = 3 then true_alpha = .69;
 if rel_items = 1  then true_alpha = .90;

if true_alpha = .33 then a_3pl = ta_3pl[1:3];
if true_alpha = .33 then b_3pl = tb_3pl[1:3];
if true_alpha = .33 then c_3pl = tc_3pl[1:3];
if true_alpha= .54 then a_3pl = ta_3pl[1:6];
if true_alpha= .54 then b_3pl = tb_3pl[1:6];
if true_alpha= .54 then c_3pl = tc_3pl[1:6];
if true_alpha = .69 then a_3pl = ta_3pl[1:11];
if true_alpha = .69 then b_3pl = tb_3pl[1:11];
if true_alpha = .69 then c_3pl = tc_3pl[1:11];
if true_alpha = .90 then a_3pl = ta_3pl[1:50];
if true_alpha = .90 then b_3pl = tb_3pl[1:50];
if true_alpha = .90 then c_3pl = tc_3pl[1:50];
items = nrow(A_3PL);
do njs_cond = 1 to 1;   * average sample size in study;
  if njs_cond = 1 then njs = 10; * actual value 10 changed to check rxx;
  *if njs_cond = 1 then njs = 50;
  *if njs_cond = 3 then njs = 100;
  *if njs_cond = 4 then njs = 500;
 *if njs_cond = 1 then njs = 1500;

do k_cond = 1 to 1;* N of studies in each meta-analysis;

 if k_cond = 1 then n_studies =  15; * actual value 15 changed to check rxx;
 *if k_cond = 2 then n_studies =  50;
 *if k_cond = 1 then n_studies = 100;
 *if k_cond = 4 then n_studies = 150;
* 30, April 2005 Jeanine added index for nr;
do Num_alpha = 1 to 5; *i reliabilities in each of the k journals;

if Num_alpha =1 then nr =1;
if Num_alpha =2 then nr = 2;
if Num_alpha =3 then nr = 3;
if Num_alpha =4 then nr = 10;
if Num_alpha =5 then nr = 50;
* +------------------------------------------------------------------------------+
    Initialize counters
 5, June 2005 Note: only weighted Z alpha is needed
    Columns is Z-alpha
    Rows are: ignoring dep alpha,mean alpha per study,median alpha per study,
    random alpha per study-- not sure about HLM level 2 alpha
```

Appendix A (continued) SAS Code for Monte Carlo Simulation

```
  +------------------------------------------------------------------------------+;

  * Mean Values;
  Means = J(4,1,0);

  * Confidence Band Coverage;
  InBand = J(4,1,0);

  * Confidence Band Width;
  WideBand = J(4,1,0);

sumrxx= J(4,1,0);

rmse=J(4,1,0);

bias = J(4,1,0);

  nsamples=0;

seed1 = round(1000000*ranuni(0));

do rep=1 to replicat;           * This starts the big do loop;

rep_vec = J(n_studies#NR,1,rep);
if rep =1 & njs_cond =1 & k_cond =1  & num_alpha=1 & rel_items = 1 then do; * add reli loop;
      create ICCout3 from rep_vec[colname = 'meta'];
      append from rep_vec;
end;
if rep >1 | njs_cond >1 | k_cond >1 | num_alpha>1 |rel_items >1 then do;  * add reli loop;
      setout ICCout3;
      append from rep_vec;
end;

  *do study = 1 to n_studies;  * Inner loop for primary studies;
```

Appendix A (continued) SAS Code for Monte Carlo Simulation

```
     * randomly generate a sample size for each study;
*do simulee = 1 to n1; * Number of examinees to generate;
  seed1=round(100000000*ranuni(0));
 * idn2 = simulee;
**+-----------------------------------------------------------+
29, April 2005 added to Generate a variance between and variance
  within to simulate intra class corr between reliabilities.
+-------------------------------------------------------------+;
  do studies = 1 to n_studies;
      mean_theta = rannor(seed1);
      mean_theta = mean_theta#sqrt(.0005);   *since this value varies I'll just type it in;

do numrel = 1 to nr;
      n1=rannor(0)#(.20#njs) + njs;
    n1=round(n1);
     if n1<4 then n1=4;
      *n1=njs;
      do simulee = 1 to n1;
                  theta= rannor(seed1);
                  theta = theta#sqrt(1)+ mean_theta;


* +----------------------------------------------------+
29, Aril 2005 -only one administration needed for alpha
  Administer the test twice to each examinee: to allow
   both Cronbach alpha and test-retest estimates
  +----------------------------------------------------+;
      run IRT3PL (THETA, A_3PL, B_3PL, C_3PL, SEED1, True_P, SCORE3PL);
      *run IRT3PL (THETA, A_3PL, B_3PL, C_3PL, SEED1, True_P, SCORE2);

* +-------------------------------------------------------------+
     Build matrix of scores for examinees
  +-------------------------------------------------------------+;
      if simulee = 1 then out3pl = score3pl;
      if simulee > 1 then out3pl = out3pl//score3pl;
       end;

 * +-------------------------------+
     Computation of Cronbach Alpha
  +-------------------------------+;

      mu1 = J(items,1,0);* Changed this!;
      var = J(1,items,0);* Changed this!;

         do k = 1 to items;
            do i=1 to n1;
          mu1[k,1] = mu1[k,1] + out3pl[i,k];
        end; * end n1;
        var[1,k]=(mu1[k,1]/n1)*(1 - mu1[k,1]/n1); * var of items;
            * print mu1 var;
      sumvar=0;
      do k = 1 to items;* sum of the item variances

        sumvar = sumvar + var[1,k];
      end; *end items;
            * print sumvar;
      rowsum = J(n1,1,0);
*     rowsum2= J(n1,1,0);
      do p = 1 to n1;
       do k = 1 to items;
          rowsum[p,1]=rowsum[p,1] + out3pl[p,k];  *calculate the row sum for each examinee;
      *     rowsum2[p,1]=rowsum2[p,1] + out2[p,k];
       end; * end n1;
      end; *end items;
            *print rowsum;
      sumscore = 0;
      sumscore2 = 0;
      do p = 1 to n1;
            sumscore = sumscore + rowsum[p,1];
            sumscore2= sumscore2 + rowsum[p,1]##2;
      end; * end n1;
      vartotal= (sumscore2-(sumscore##2/n1))/(n1); *var of all examinees total score;
      * +------------------------------------------------------------+
         Be sure we have some score variance before going any further
      * +------------------------------------------------------------+;
    if vartotal > 0 then do;
            *print sumscore sumscore2 vartotal;
            rxx = (items/(items -1))*((vartotal- sumvar)/vartotal); * This is Cronbach alpha!;
*print n1 out3pl rxx;

if rxx < 0.00001 then rxx = .00001; * need to confirm rxx;
if rxx> .9999 then rxx = .9999;*Jeff change;
*print 'Check Values of rxx';
*print studies numrel rxx;
***********************************************************************************;
            *Jeanine Add code to create matrix;
if (studies = 1 & numrel = 1) then do;
 *xbartheta = mean_theta;
t_alpha_vec= true_alpha;
No_alpha_vec = nr;
njs_vec=njs;
n_studies_vec=n_studies;
study = studies;
 est_rel = numrel;

 z_rxx = log(1-abs(rxx));
rxx_vec = rxx;

sd_vec = ((2#items)/((items - 1) #(n1 - 2)))##-1;
  n_vec= n1;
 *sd_vec = sd;
 * print true_alpha rxx_vec;
end; * end  studies = 1;
if (studies > 1 | numrel > 1) then do;
```

```
 *xbartheta = xbartheta//mean_theta;
t_alpha_vec= t_alpha_vec//true_alpha;
No_alpha_vec = No_alpha_vec //nr;
njs_vec=njs_vec//njs;
n_studies_vec=n_studies_vec//n_studies;
study = study//studies;
 est_rel = est_rel//numrel;
 z_rxx = z_rxx//log(1-abs(rxx));
rxx_vec = rxx_vec//rxx;
n_vec= n_vec//n1;
 sd_vec = sd_vec//((2#items)/((items - 1) #(n1 - 2)))##-1;
end; *end studies >1;
*print mean_theta studies numrel rxx;
*print studies;
*print rxx;
end;  *end vartotal;
if vartotal = 0  then Numrel= numrel-1;
              end;  * end big n_studies loop;
               end;  * end big nr loop;
ri_by_k= J(nr,n_studies,0);
N_vec_mtx=J(nr,n_studies,0);
*print 'first';
*print ri_by_k;
do v = 1 to n_studies;
do i =1 to nr;
w =nr#(v-1)+ i;
ri_by_k[i,v]=rxx_vec[w,1];
 N_vec_mtx[i,v]=n_vec[w,1];

          end; * end N_studies above;
           end; * end nr above;
*print N_vec_mtx;
*print rxx_vec;* z_rxx;

if rep =1 & njs_cond =1 & k_cond =1  & num_alpha=1 & rel_items = 1 then do; * add reli loop;
```

```
      create ICCout1 from study[colname ='study'];
            append from study;
      create ICCout2 from sd_vec[colname = 'weightv'];
            append from sd_vec;
      create ICCout4 from z_rxx[colname = 'rxx_vec'];
            append from z_rxx;
      create ICCout5 from t_alpha_vec[colname = 'true_alpha'];
            append from t_alpha_vec;
      create ICCout6 from No_alpha_vec[colname = 'num_rel'];
            append from No_alpha_vec;
      create ICCout7 from njs_vec[colname = 'njs'];
            append from njs_vec;
      create ICCout8 from N_studies_vec[colname = 'N_studies'];
            append from N_studies_vec;
end;

if rep >1 | njs_cond >1 |k_cond >1 | num_alpha>1 | rel_items > 1 then do; * add reli loop;
      setout ICCout1;
            append from study;
      setout ICCout2;
            append from sd_vec;
      setout ICCout4;
            append from z_rxx;
      setout ICCout5;
            append from t_alpha_vec;
      setout ICCout6;
            append from No_alpha_vec;
      setout ICCout7;
            append from njs_vec;
      setout ICCout8;
            append from N_studies_vec;
end;


*
```

```
            *------------------------------------+
          * calculate sample standard deviation and SEM;
*----------------------------------------------------------------
       ss1 = (J(1,n1,1)*(rowsum##2)) - ((J(1,n1,1)*rowsum)##2/n1);
   *   sd = sqrt(ss1/(n1-1));
  *________________________________________________________________;

       * collect KR21, rxx, SEM, n and sd in vectors;
           *      if study = 1 then KR_vec= KR;
                   *          if study >1 then KR_vec = KR_vec//KR;
            *     if study = 1 then retest_vec = rxx_retest;
            *           if study > 1 then retest_vec = retest_vec//rxx_retest;
            *     if study = 1 then rxx_vec = rxx;
            *           if study > 1 then rxx_vec = rxx_vec//rxx;
            *     if study = 1 then n_vec = n1;
            *           if study > 1 then n_vec = n_vec//n1;
            *     if study = 1 then sd_vec = sd;
            *           if study > 1 then sd_vec = sd_vec//sd;
    *end; * end the 'if vartotal > 0 then do' loop;;



*end; * end the studies loop;
* print rxx_vec retest_vec KR_vec n_vec sd_vec;
*+-------------------------------------------------
Ignore dep part-
The calc_alpha calculates a vector of alphas for all
the studies
+--------------------------------------------------;
        * compute mean reliability for the sample of studies;
            run calc_alphaVI(ri_by_k,items,N_vec_mtx,Z_W_mean,SE_Z);
            *print 'Ignore' z_w_mean;
      * +-------------------------------------------------------+
            Compute mean values, bandwidths and band coverage here
            for all studies i.e ignoring dependence
            Only using weighted Fishers z for this study
      * +--------------------------------------------------------+;
   * means[1,1] = means[1,1] + w_alpha;
    means[1,1] = means[1,1] + (1 - exp(z_w_mean));
   *wideband[1,1] = wideband[1,1] + ((W_alpha + 1.96#SE_alpha) - (W_alpha - 1.96#SE_alpha));
    wideband[1,1] = wideband[1,1] + (1 - exp(z_w_mean - 1.96#SE_z)) - (1 - exp(z_w_mean + 1.96#SE_z));

if (true_alpha > (1 - exp(z_w_mean + 1.96#SE_Z)) & true_alpha < (1 - exp(z_w_mean - 1.96#SE_Z))) then
inband[1,1] = inband[1,1] + 1;
  sumrxx[1,1]=sumrxx[1,1]+((1-exp(z_w_mean))-true_alpha)##2;
  bias[1,1] =bias[1,1]+ ((1-exp(z_w_mean)) - true_alpha);
*-------------------------------------------+
      29, April 2005 changed for J9 dis
*-------------------------------------------;
 free z_w_mean SE_z;

    * end; * End analysis for ingnoring dep;
*+-------------------------------------------------
calculating mean alpha per study part
The calc_alpha calculates a vector of alphas for all
the studies

+--------------------------------------------------;
        * compute mean reliability for the sample of studies;
            run calc_kalpha_mean(ri_by_k,items,N_vec_mtx,Z_W_mean,SE_Z);
      *     print 'mean' z_w_mean;
      * +--------------------------------------------------------+
Compute mean values, bandwidths and band coverage here
            for all studies i.e ignoring dependence
            Only using weighted Fishers z for this study
      * +--------------------------------------------------------+;
   * means[1,1] = means[1,1] + w_alpha;
    means[2,1] = means[2,1] + (1 - exp(z_w_mean));
   *wideband[1,1] = wideband[1,1] + ((W_alpha + 1.96#SE_alpha) - (W_alpha - 1.96#SE_alpha));
    wideband[2,1] = wideband[2,1] + (1 - exp(z_w_mean - 1.96#SE_z)) - (1 - exp(z_w_mean + 1.96#SE_z));
    if (true_alpha > (1 - exp(z_w_mean + 1.96#SE_Z)) & true_alpha < (1 - exp(z_w_mean - 1.96#SE_Z))) then
inband[2,1] = inband[2,1] + 1;
    sumrxx[2,1]=  sumrxx[2,1]+ ((1-exp(z_w_mean))-true_alpha)##2;
     bias[2,1] =bias[2,1]+ ((1-exp(z_w_mean)) - true_alpha);
*-------------------------------------------+
      29, April 2005 changed for J9 dis
*-------------------------------------------;
       free z_w_mean SE_z;

     *end; * End analysis for calculating one mean per study;
*+-------------------------------------------------
calculating Md alpha per study part

+--------------------------------------------------;
        * compute mean reliability for the sample of studies;
            run calc_kalpha_Med(ri_by_k,items,N_vec_mtx,Z_W_mean,SE_Z);
            *print 'median' z_w_mean;
      * +--------------------------------------------------------+
            Compute mean values, bandwidths and band coverage here
            for all studies i.e ignoring dependence
            Only using weighted Fishers z for this study
      * +--------------------------------------------------------+;
   * means[1,1] = means[1,1] + w_alpha;
    means[3,1] = means[3,1] + (1 - exp(z_w_mean));
   *wideband[1,1] = wideband[1,1] + ((W_alpha + 1.96#SE_alpha) - (W_alpha - 1.96#SE_alpha));
    wideband[3,1] = wideband[3,1] + (1 - exp(z_w_mean - 1.96#SE_z)) - (1 - exp(z_w_mean + 1.96#SE_z));
    if (true_alpha > (1 - exp(z_w_mean + 1.96#SE_Z)) & true_alpha < (1 - exp(z_w_mean - 1.96#SE_Z))) then
inband[3,1] = inband[3,1] + 1;
    sumrxx[3,1]=sumrxx[3,1]+ ((1-exp(z_w_mean))-true_alpha)##2;
     bias[3,1] =bias[3,1]+ ((1-exp(z_w_mean)) - true_alpha);
*-------------------------------------------+
      29, April 2005 changed for J9 dis
*-------------------------------------------;
free z_w_mean SE_z;
```

```
     *end; * End analysis for calvulating one median per study;

*+-------------------------------------------------
calculating rand alpha per study part
+--------------------------------------------------;
        * compute mean reliability for the sample of studies;
            run calc_kalpha_rand(ri_by_k,items,N_vec_mtx,Z_W_mean,SE_Z);
      * +-------------------------------------------------------+
            Compute mean values, bandwidths and band coverage here
            for all studies i.e ignoring dependence
            Only using weighted Fishers z for this study
      * +-------------------------------------------------------+;
   * means[1,1] = means[1,1] + w_alpha;
    means[4,1] = means[4,1] + (1 - exp(z_w_mean));
   *wideband[1,1] = wideband[1,1] + ((W_alpha + 1.96#SE_alpha) - (W_alpha - 1.96#SE_alpha));
    wideband[4,1] = wideband[4,1] + (1 - exp(z_w_mean - 1.96#SE_z)) - (1 - exp(z_w_mean + 1.96#SE_z));

   if (true_alpha > (1 - exp(z_w_mean + 1.96#SE_Z)) & true_alpha < (1 - exp(z_w_mean - 1.96#SE_Z))) then
inband[4,1] = inband[4,1] + 1;
     sumrxx[4,1]=sumrxx[4,1]+  ((1-exp(z_w_mean))-true_alpha)##2;
   bias[4,1] = bias[4,1]+((1-exp(z_w_mean)) - true_alpha);
*-------------------------------------------+
      29, April 2005 changed for J9 dis
*-------------------------------------------;
free z_w_mean SE_z;

     *end; * End analysis for calculating one median per study;

  nsamples=nsamples+1;
 * print means;
end;   *end the big loop rep end;

*print means;
* +-----------------------+
   Convert sums into means
```

Appendix A (continued) SAS Code for Monte Carlo Simulation

```
  +-----------------------+;
do row = 1 to 4;
    if means[row,1] ^= . then means[row,1] = means[row,1]/nsamples;
    if InBand[row,1] ^= . then InBand[row,1] = InBand[row,1]/nsamples;
    if WideBand[row,1] ^= . then WideBand[row,1] = WideBand[row,1]/nsamples;
      if Bias[row,1]^=. then Bias[row,1] = Bias[row,1]/nsamples;
      if sumrxx[row,1]^= . then sumrxx[row,1] = sumrxx[row,1]/nsamples;
      rmse[row,1]= sqrt(sumrxx[row,1]);
 end; * end row;

*print 'Reliability Generalization';
 label1='Violation';
 label2 = 'Mean';
 label3 = 'Median';
 label4 = 'Random';
 labels = label1//label2//label3//label4;
print labels icc true_alpha nr njs  n_studies Bias RMSE InBand WideBand means nsamples;

end; * end the k_cond loop;
end; * end the njs_cond loop;
end; * end the rel_items loop;
end; * end the num_alpha loop;


data allout;
 merge iccout1 iccout2 iccout3 iccout4 iccout5 iccout6 iccout7 iccout8;*proc print data=allout;
*proc print data =allout;
*proc means data =allout;
*var rxx_vec;
proc datasets;
delete iccout1 iccout2 iccout3 iccout4 iccout5 iccout6 iccout7 iccout8;
proc sort data = allout;
 by meta true_alpha num_rel njs N_studies;
Proc mixed noclprint covtest noitprint noinfo;
by meta true_alpha num_rel njs N_studies;
```

Appendix A (continued) SAS Code for Monte Carlo Simulation

```
class study;
weight weightv;
model rxx_vec= /solution CL;
random intercept/sub =study;
ods output solutionF= rgsim
(keep= meta true_alpha num_rel njs N_studies estimate lower upper);
ods output FitStatistics= rgCI;

ods listing close; *after that;

run;

*title 'proc mixed random';
*proc print data = rgsim;
ODS LISTING; *Jeff change;
proc sort data =rgsim;
by  true_alpha num_rel njs N_studies;
data rgsim2;
set rgsim;
*proc transpose data = rgsim out= rgsim2;
*PROC CONTENTS DATA =RGSIM2;
dm 'log; clear;' continue;
proc means noprint data = rgsim2;
by true_alpha num_rel njs N_studies;
var estimate;
OUTPUT OUT = MIX mean= Zrxx_est;

*proc contents data= mix;
data mix_trans;
MERGE mix RGSIM2;
BY true_alpha num_rel njs N_studies;
orig_r_mean= 1- exp(zrxx_est); *++++++transforms zmean back to alpha this is the estimate of
true_alpha+++++;
 *wideband[2,1] = wideband[2,1] + (1 - exp(z_w_mean - 1.96#SE_z)) - (1 - exp(z_w_mean + 1.96#SE_z));
*untrans = upper-lower;
```

```
upper_rxx = 1-exp(upper);*Jeff change;
lower_rxx = 1-exp(lower);
wideband=abs(upper_rxx-lower_rxx);

if true_alpha < lower_rxx & true_alpha> upper_rxx then inband =1;
else inband =0;
bias = orig_r_mean-true_alpha;
*DATA RMSE_CAL;
*SET MIX_TRANS;
sumrxx =((1- exp(estimate))-true_alpha)**2; *Jeff change again;

*proc print;
*var upper lower wideband wideband3 upper_rxx lower_rxx;
*PROC CONTENTS DATA = MIX_TRANS;
*proc print data=mix_trans;
*run;
proc means noprint data = mix_trans;
by true_alpha num_rel njs N_studies;
var wideband;
output out = rgmeans1 mean= av_wideband;


proc means noprint data = mix_trans;
by true_alpha num_rel njs N_studies;
var inband;
output out = rgmeans2 mean = av_inband;

proc UNIVARIATE noprint data = mix_trans;
by true_alpha num_rel njs N_studies;
var sumrxx;
output out = rgmeans3 mean = rxx_sum;
data rgmeans;
```

Appendix A (continued) SAS Code for Monte Carlo Simulation

```
icc= .01;
merge mix_trans rgmeans1 rgmeans2 rgmeans3;
by true_alpha num_rel njs N_studies;
*rxx_sum= (sum(sumrxx));
Rmse= sqrt(rxx_sum); *Jeff change;

if first.njs or first.n_studies; * Jeff change;

proc print data = rgmeans;
by num_rel njs;

var icc true_alpha num_rel njs N_studies bias rmse av_inband av_wideband orig_r_mean;
run;
```

About the Author

Jeanine Romano received a bachelor degree (BS) in Mathematics education in 1994 from the University of South Florida and a Masters degree (MA) in Mathematics education in 1996. She had worked as an instructor and the Coordinator of Institutional Research and Assessment at The University of Tampa. At the University of Tampa she taught lower level mathematics courses and statistics. In addition, she has taught undergraduate measurement course both face to face and on line. Her research has been nominated for the Florida Educational Research Association distinguished paper five times and had won the award both in 2004 and in 2006. Her research was recently recognized as Best Paper for the 2007 Florida Association of Institutional Research Conference.