7-10-2003

A Comparison of Meta-Analytic Approaches to the Analysis of Reliability Estimates

Denise Corinne Mason
University of South Florida

A Comparison of Meta-Analytic Approaches to the Analysis of Reliability Estimates

by

Denise Corinne Mason

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Psychology
Department of Psychology
College of Arts Sciences
University of South Florida

Major Professor: Michael Brannick, Ph. D.
Walter Borman, Ph.D
Carnot Nelson, Ph.D
Billy N. Kinder, Ph.D
Joel Thompson Ph.D

Date of Approval:
July 10, 2003

Keywords: statistics, testing methodology, validity, r to z transformation, moderators

Dedication

To Dr. Mike Brannick, without whom I never would have finished this degree, thank you always for believing in me.

To my Mom and Dad, you both supported me through this in your own special ways, and now we have the first Ph.D. in the family. Thank you both for teaching me to dream big, work hard and never give up.

To my sis, Rhonda, you are the best editor in the world and you are permanently employed as my best friend and my editorial staff, thank you for the details.

To my friends, Ali, Sharon, John, Don, Karen, Dr. Ed and all the others who have cheered me on through this process, you were there for me when I needed encouragement. All of you are honorary Ph.D.'s of Motivation, in my life. Thank you!

And to Reynald, whose genius and love shine forth through this paper in significant ways, thank you for keeping me on track, mentoring me and standing beside me all the way!

Table of Contents

List of Tables

List of Figures

A Comparison of Meta-analytic Approaches to the Analysis of Reliability Estimates

Denise Corinne Mason

ABSTRACT

In the last few years, several studies have attempted to meta-analyze reliability estimates. The initial study, to outline a methodology for meta-analyzing reliability coefficients, was published by Vacha-Haase in 1998. Vacha-Haase used a very basic meta-analytic model to find a mean effect size (reliability) across studies. There are two main reasons for meta-analyzing reliability coefficients. First, recent research has shown that many studies fail to report the appropriate reliability for the measure and population of the actual study (Vacha-Haase, Ness, Nilsson and Reetz, 1999; Whittington, 1998; Yin and Fan, 2000). Second, very little research has been published describing the way reliabilities for the same measure vary according to moderators such as time, form length, population differences in trait variability and others. Vacha-Haase (1998) proposed meta-analysis, as a method by which the impact of moderators may become better understood.

Although other researchers have followed the Vacha-Haase example and meta-analyzed the reliabilities for several measures, little has been written about the best methodology to use for such analysis. Reliabilities are much larger on average than are validities, and thus tend to show greater skew in their sampling distributions.

This study took a closer look at the methodology with which reliability can be meta-analyzed. Specifically, a Monte Carlo study was run so that population characteristics were known. This provided a unique ability to test how well each of three methods estimates the true population characteristics. The three methods studied were the Vacha-Haase method as outlined in her 1998 article, the well-known Hunter and Schmidt "bare bones method" (1990) and the random-effects version of Hedges's method as described by Lipsey and Wilson (2001). The methods differ both in how they estimate the random-effects variance component (or in one case, whether the random-effects variance component is estimated at all) and in how they treat moderator variables. Results showed which of these methods is best applied to reliability meta-analysis. A combination of the Hunter and Schmidt (1999) method and weighted least squares regression is proposed.

Introduction

For years a debate has raged concerning the utility of the social sciences in light of an apparent lack of clarity around research findings. (Hunter and Schmidt 1990; Lipsey and Wilson, 2001; Rosenthal, 1987). This debate seems to be fueled by the habit of behavioral and social scientists to consistently call for more research in the discussion and concluding remarks of published studies.

In an effort to quell the criticisms leveled at the social sciences, various methods for aggregating data across studies have been developed in the hope that aggregate data analysis would provide the social sciences more surety in drawing conclusions. Many of the earliest methods of aggregation were based on literature reviews. Conclusions were drawn based on the reviewers' overall perceptions of what each study added to the current knowledge in the area. However, such qualitative analyses left many unanswered questions because of the potential for bias.

A Brief History of Meta-Analysis

In the late 1960's and early 1970's one of the major debates within the behavioral and social sciences concerned the effectiveness of therapy in clinical psychology. Reviews of the literature had left many wondering whether clinical therapy was effective. Gene Glass (1976) presented what he called "meta-analysis" as a way to combine the results of multiple studies in a quantitative way. He and a colleague analyzed over 400 studies designed to assess the effectiveness of psychotherapy. He was able to show that, on average, across a large number of studies, therapy made a significant difference in the client outcomes.

Glass (1976) provided this example to show how meta-analysis could be used to compute an average effect size across studies. Glass also demonstrated that such averaged effect sizes could be used to find conclusions among opposing findings. Prior to meta-analysis, most methods for summarizing studies failed to incorporate the effect-size statistics and instead simply summarized the findings on a categorical basis (i.e., significant vs. not). An effect-size statistic is the index used to represent study findings in direction and magnitude (Lipsey and Wilson, 2001). Meta-analysis is essentially the survey research method by which the effect size of the research studies is surveyed, weighted and compared.

Glass's meta-analytic method caught the eye of many psychologists and remains well cited in the social sciences. Other meta-analytic pioneers include Rosenthal (1987), who studied experimenter expectancy effects, and Schmidt and Hunter (1977), who studied employment testing. All such studies have now been labeled "meta-analysis" but each method has its own specific idiosyncrasies.

Within the Industrial and Organizational literature, the Schmidt and Hunter (1977) (later Hunter and Schmidt, 1990) method of meta-analysis is probably the most well-cited and -used model for analyzing study results. Of particular interest in this field has been the study of the validity of personnel tests. Schmidt and Hunter (1977) introduced the concept of "validity generalization." They presented the theoretical position that in the test validation context, test validity is a constant as long as all the following elements are equivalent: (a) job family (b) type of test and (c) criterion of overall job performance. They then built a step-by-step meta-analytic method based on that theoretical assumption. Their meta-analytic approach became known as validity generalization.

The popularity of the Schmidt and Hunter approach is apparent, as the majority of published meta-analyses with Industrial and Organizational psychology have focused on validity generalization (Hall, 2000). However, there is potentially a difficulty in using this method for reliability because it was developed specifically for validity. It is apparent that validity is always impacted by reliability, but what subtle difference in methodology might there be when looking at the relationship from the reliability perspective alone? Although the Schmidt and Hunter (1977) and Hunter and Schmidt (1990) took the reliability of the test into consideration as one of the "artifacts" in the study, they treated reliability reporting as a secondary consideration.

To be fair, Hunter and Schmidt (1990) did include a method to estimate reliability using hypothetical distributions in their procedure when reliability is not reported. However, neither Schmidt and Hunter (1977) nor Hunter and Schmidt (1990) focused directly on the estimation of reliability across studies. Therefore, the degree to which their procedures apply to reliability estimates rather than validity estimates is something of an open question.

Recent reviewers of the meta-analysis of reliability data by Vacha-Haase (1998) and others (Vacha-Haase, Ness, Nilsson and Reetz, 1999; Yin and Fan, 2000; Whittington, 1998) have shown that published studies rarely incorporate the correct reliability estimates. Vacha-Haase et al. (1999) noted that as many as half of all studies fail to report the reliability estimates based on that study’s data. Such omissions occur despite the American Psychological Association's (APA, 1994) encouragement in the publication guidelines to report effect size, reliability and related statistics for *each* study.

Because reliability is not reported in many studies and because reliability directly impacts validity, validity estimates for individual studies may be erroneous or misleading to an unknown degree. One obvious means of combating the problem is to report the reliability for the local study. Another less obvious means is to estimate the reliability of the study results after the fact from data in other studies. Note that even if the local reliability is estimated, the accuracy of the estimate will depend upon the sample size of the local study. Small samples provide estimates with relatively large sampling variances. Vacha-Haase (1998) recognized this and suggested a meta-analytic approach to assessing reliability within multiple studies. Although this approach is creative, the application of meta-analytic methods to reliability estimates may prove troublesome. The goal of this project was to investigate the application of meta-analysis methods to reliability data in order to provide some recommendations about which techniques appear best suited to the analysis. The paper is organized by the following steps:

1. Review the basics of reliability,
2. Describe how inappropriate reliability estimates can impact the current status of the literature,
3. Review the current status of the meta-analysis of reliability estimates,
4. Compare estimates from current methods analysis to known parameters in order to make recommendations about which techniques appear best under what conditions.

A Review of Reliability

In the early 1900's Spearman introduced the Classical Measurement Theory. In his theory he defined reliability as "the consistency with which individuals are rank ordered by measurement across parallel test forms, repeated measures or other estimates of consistency in measurement" (Spearman, 1910, p. 272).

Since that time, researchers within the Industrial and Organizational Psychology literature have created hundreds of assessments. Researchers have usually estimated the reliability across the studies using the following recognizable measures:

Cronbach's Alpha. Cronbach's Alpha is based on a single administration of the test. Cronbach's Alpha estimates the correlation between 'randomly parallel' tests or hypothetical sets of items 'just like these' (Nunnally and Berstein, 1994). Cronbach's Alpha is the most frequently reported reliability statistic, but it is difficult to meta-analyze because its sampling distribution is unknown.

Kuder-Richardson's Formula. Kuder-Richardson's formula is based on a single administration of the test and is used specifically with dichotomously scored data.

Split-half reliability eoefficient. Split-half reliability coefficient is based on a single administration of the test. The Split-half reliability coefficient is the single-administration analog to alternate forms reliability estimates. According to the split-half method, reliability is estimated by computing the correlation between two subsets of the overall measure.

Test-retest. Test-retest is the comparison of scores reusing the same measure.

Alternate Forms method. Alternate Forms method is the comparison of the scores based on equivalent measures (Nunnally, 1978). Of course, two different forms can also be given at two different times and compared this type of correlation is sometimes referred to as a coefficient of stability and equivalence (Cronbach and Gleser, 1964).

Although all of these forms of reliability have been used in the literature for over 30 years, reliability still remains an elusive concept to many. This may be due in part to the multiple ways in which it is calculated. However, a lack of understanding of reliability may be part of the reason why it is under- or mis-reported.

The Debate Over the Meaning of Reliability

Although the estimation of reliability may take on many forms, the underlying assumption in all of these formulas is that reliability is based on the scores obtained from the measures and not on the measures themselves (Thompson and Vacha-Haase, 2000). Despite the statistical assumption however, the psychometric translation of the concept of reliability seems to have undergone an interesting shift in meaning. As Sawilowsky (2000) noted, "reliability has become associated with the measure or test itself and its basis in the sample scores seems to have become less clear". This lack of clarity has led authors of the current literature to debate the meaning and subsequently the reporting of reliability in the literature.

Thompson and Vacha-Haase (2000) argued that endemic confusion surrounding the meaning of reliability has created false confidence in reports of a measure's reliability. As an example, they cite the number of times authors directly report reliability coefficients from the test manual as if they were a number that traveled with the test despite the population. Thompson and Vacha-Haase (2000) concluded that many authors misunderstand the impact that the lack of sample-based reliability reports has on other results like validity.

Low reliabilities lessen statistical power, increase error and attenuate effect sizes. This can lead to less correct interpretations of the validity estimates. When misreported reliabilities are translated to the multitude of meta-analytic studies that combine the validity estimates across studies, this impact is compounded (Thompson and Vacha-Haase, 2000). As previously mentioned, much of the meta-analytic work within the Industrial and Organizational Psychology literature has focused on validity generalization. It is possible to conclude that some of the interpretations made from these meta-analyses are not completely accurate due to issues surrounding reliability. Some meta-analytic methods attempt to address this issue by creating a hypothetical distribution of estimated reliabilities (Hunter and Schmidt, 1990) however even these distributions are not as perfectly correct as the actual reliability statistic would be. Incorrect assumptions of validity may also lead to the use of tests and measures in populations where they may not be appropriate or where additional factors may warrant consideration.

The discovery of this confusion over the meaning and reporting of reliability could be a huge wake-up call for the research community. If reliability estimates are largely missing or falsely reported in the literature due to a basic misunderstanding of the relationship between reliability and the actual test scores, what can be done to correct the misunderstanding and to correct assumptions based on erroneous reliability reports?

A Meta-Analytic Approach to Analyzing Reliability

Vacha-Haase et. al. (2000) have coined the term “reliability induction” to refer to the practice of explicitly referencing the reliability coefficients from prior reports as the sole warrant for presuming the score integrity of entirely new data. They argue that this is what most researchers seem to presume and why they fail to calculate and report subsequent sample-based reliability.

The most ideal solution to this issue would be to have every study report the estimated reliability based on the actual sample. However, since researchers cannot recalculate reliability for all the studies in the literature on any particular test or measure, they must find another solution. Vacha-Haase (1998) has proposed a meta-analytic procedure that helps to estimate the reliability across samples and to evaluate the additional factors that may contribute to the variability in the reliability estimate.

Vacha-Haase (1998) called this approach "reliability generalization". Using this method, she attempts to (a) examine how score reliability varies across studies (b) estimate the typical reliability of scores for a given test across studies, (c) examine the amount of variability in reliability coefficients for specific measures, and (d) identify some of the sources of variability. The reliability score's variability across studies is equivalent to the estimated population variance. The typical reliability score is analogous to the mean effect size from a meta-analysis of reliabilities. To look for the amount of variability in the actual reliability coefficients that would be attributed to a random effect variance (ie. not sampling error or moderator variance) there would need to be an estimate of a random effects component. This is something that is discussed in more detail later in this paper in the description of the Lipsey and Wilson method. Finally, to identify sources of variability, one would need to identify and analyze for moderators. It can be thus inferred, that the ideal meta-analytic technique according to Vacha-Haase would be able to provide a mean effect size estimate, provide an estimate of the variance around that mean, account for the expected variation within the mean do to random effects related to true score error and provide a reasonable way to deal with moderator analysis.

As a brief side note, although Vacha-Haase used the phrase "reliability generalization," the introduction of new jargon seems unnecessary; therefore this author will instead refer to this procedure as the meta-analysis of reliability.

As previously mentioned, reliability estimates are often under-reported, or even mis-reported as the reliability from testing manuals. In the absence of local reliability estimates, researchers need a way to determine a range of reliability for a measure across studies and they need some identification of the factors that moderate the change in reliability estimates across different studies. The more that reliability can be understood as a function of local conditions (such as the variability of the true scores, the type of reliability estimate, and so forth), the better researchers can estimate the true reliability within a study.

This same line of thinking may also have a profound effect on the way in which researchers understand reliability and its meaning. If, for example, test manuals could show a range of reliabilities across various situations and contexts for a test and discuss why an accurate estimate must be based on the actual population that the researcher is using (rather than the typical reliability estimate based on the validation study alone), maybe the importance of the reliability estimate would be more clear. Perhaps seeing the ranges and understanding -- in a more visible way -- that reliability changes across studies, may help to alleviate some of the misunderstanding around reliability as outlined by Thompson and Vacha-Haase (2000).

For all these reasons, the concept of meta-analyzing reliabilities clearly makes sense. However, the I/O Psychology literature has been largely devoid of such meta-analyses until the late 1990s. This may explain why the Vacha-Haase (1998) article has been quickly followed by several similar analyses of various tests and measures (Yin and Fan, 2000; Viswesvaran and Ones, 2000; Caruso et. al., 2001).

The meta-analysis of reliability is a whole new field for meta-analytic techniques. Vacha-Haase (1998) stated that she was modeling her technique after the Hunter and Schmidt (1990) meta-analysis method. However, on closer investigation, her method does not exactly match that of Hunter and Schmidt (1990). In essence, she has created a revised method and other researchers have followed her lead. Yet the method is still somewhat underdeveloped.

A logical next step for the literature when addressing the meta-analysis of reliability should be to concentrate on the methodology that can produce the best estimates of population values, as well as moderators. In an effort to highlight the current state of the literature, an explanation and comparison of the Hunter and Schmidt method and the Vacha-Haase revision are considered next.

Differences Between Vacha-Haase and Schmidt and Hunter

Vacha-Haase (1998) recommended a method to combine reliabilities based on the Schmidt and Hunter validity generalization model (Schmidt and Hunter 1977, Hunter and Schmidt 1990). The studies that have followed repeated this example (Caruso et. al., 2001; Yin and Fan, 2000).

Vacha-Haase most likely used this method as a model because the Schmidt and Hunter method is one of the most frequently cited meta-analysis methods in the Industrial and Organizational Psychology literature (Hall and Brannick, 2002). However, the Hunter and Schmidt (1990) method and the Vacha-Haase (1998) method for analyzing reliabilities contain some critical differences. These differences are so great as to suggest two different techniques and possibly significantly different outcomes. To highlight these differences, a brief review of the revised Hunter and Schmidt (1990) method followed by a description of, and comparison, to the Vacha-Haase (1998) method is outlined next.

Schmidt and Hunter method of meta-analysis. Schmidt and Hunter (1977) proposed a meta-analysis method developed specifically to support their theory that in personnel selection testing there is "one true validity" per any specific job family. They proposed that any variance in validity estimates across studies within a job family was due to sampling error and other 'artifacts' (Hunter & Schmidt, 1990; Schmidt & Hunter, 1977). They provided an example showing that error variance in a small sample size is enough to draw erroneous conclusions about moderator effects and about outcomes in general. As alternatives to significance testing, they recommended using confidence intervals in single studies and meta-analytic procedures where multiple studies are available (Hunter and Schmidt, 1990).

As does any method, the Schmidt and Hunter method has gathered some criticisms. For example, some researchers have disagreed with the criteria used to determine which studies are included in a validity generalization meta-analysis. In any meta-analysis, analysts decide which studies are included according to how well the studies fit certain inclusion limits. Some researchers believe that the Schmidt and Hunter method makes too many assumptions about how similar the predictor-criterion relationships are in personnel testing (Algera, Jansen, Roe, and Vijn, 1984). Despite criticisms, the Schmidt and Hunter method seems to be the most frequently occurring method used in the Industrial and Organizational Psychology literature (Hall and Brannick, 2002) and has been used in repeated meta-analytic studies. In the most basic outline of the Hunter and Schmidt (1990) method there are five basic steps involved in the meta-analytic process:

1. **Calculate the desired descriptive statistic** for each study available, and average the statistic across studies.
2. **Calculate the variance of the statistic** across studies.
3. **Correct the variance** by subtracting the amount attributed to sampling error. This is done by estimating the amount of variance due to sampling error (σ_e^2) with the formula:

   $$\sigma_{\mathrm{e}}^2 = \left(1-\bar{r}^2\right)^2 / \left(\bar{N}-1\right)$$

   where the $\bar{r}^2$ is a weighted mean of observed correlation values and $\bar{N}$ is the mean number of participants per study.
4. **Correct the mean and variance for study artifacts** other than sampling error.
5. **Compare the corrected standard deviation to the mean** to assess the size of the potential variation in results across studies.

6. **Consider Moderator Variables**. The moderator analysis proposed by Hunter and Schmidt (1990) includes a series of meta-analytic procedures, where validities are divided into groups based on moderators and then each group is individually meta-analyzed.

Vacha-Haase, in her 1998 article, used observably different steps to conduct a meta-analysis of reliabilities for the Bem Sex Role Inventory (BSRI).

The Vacha-Haase method. Vacha-Haase (1999) employed more of a three-step model of meta-**Characterize typical reliability and variability of score** analysis. A basic outline of these steps is as follows:

1. **Reliability coefficients expressed in squared metrics**. She used a box-and-whisker plot to represent these results.

2. **Develop a coding system to code features of the study that are predicted to impact reliability.** Vacha-Haase used type of reliability coefficient, long vs. short forms, gender of participant, article type, language the test was conducted in and sample type (e.g. student vs. non-student) and finally response format.

3. **Perform ordinary least squares regression analysis to explore how well the coded study features predict variations in the reliability coefficients.** She uses this analysis to identify the differential influences of various sources of measurement error in order to better predict what the reliability coefficient would look like in a new sample. Vacha-Haase

presented these results in a table with the R^2's and beta weights for each predictor variable.

Next, the step-by-step differences between the two methods will be explored and an explanation will be offered as to how these discrepancies may produce incongruous outcomes.

Highlighting the differences. In each of the following steps, the Hunter and Schmidt method is outlined first and then compared with the Vacha-Haase approach.

Step 1: Desired descriptive statistic and average of that statistic across studies.

With the Hunter and Schmidt (1990) meta-analytic approach to validity, the effect-size statistic is the validity coefficient. The validity coefficients across research studies are the unit of interest and the average is displayed as a mean validity coefficient. This mean is important because it represents, in the Hunter and Schmidt theory (1990), the true validity of the test regardless of the situation in which the test is given.

When meta-analyzing reliability, as in the Vacha-Haase method, the reliability coefficient (r_{xx}) is the effect-size statistic used to average across studies. The reliability coefficient is represented as a correlation coefficient, which has a range from –1 to +1. As with validity coefficients, mean reliability can be calculated. Vacha-Haase computed a unit-weighted average rather than a sample-size-weighted average. By choosing not to weight by sample size, Vacha-Haase is departing from a practice that most meta-analytic techniques incorporate (Hunter & Schmidt, 1990; Lipsey and Wilson, 2001; Rosenthal, 1984).

If studies are randomly drawn from a population, weighting them by a function of their precision will result in an estimate of the mean that has a smaller sampling variance than what is obtained by unit weights (e.g., Hedges, 1985; Raju & Drasgow, in press). Because the precision weighted mean should have a smaller standard error than the unit weighted mean, the precision weighted mean is generally preferred (Lipsey & Wilson, 2001). However, if the sample size is correlated with the effect size, the use of the precision-weighted mean can result in a biased estimate of the meta-analytic mean (Overton, 1998). Vacha-Haase (1998) found that sample size was correlated with effect size in a meta-analysis of the Bem Sex Role Inventory, at least for the Female scale.

Hunter and Schmidt (1990) used the sample size as the weight. However, because they use *r* rather than *z* in the analysis, the weight is not equal to the inverse of the sampling variance (Lipsey & Wilson, 2001; Raju & Drasgow, in press). (For *z*, the inverse variance weight is N-3; for *r*, the inverse variance weight is $\frac{N-1}{(1-r^2)^2}$). Raju and Drasgow (in press) described the Hunter and Schmidt (1990) method, as based on the method of moments, and the inverse variance weights (Lipsey & Wilson, 2001) as based on the method of maximum likelihood. The inverse variance weights have the desirable property of having the minimum sampling variance of any estimator of the mean (Hedges, 1985; Lipsey & Wilson, 2001; Raju & Drasgow, in press).

Step 2: Calculate the variance of the statistic across studies.

Hunter and Schmidt (1990) pointed out that if the population correlation is assumed to be constant over studies, then the best estimate of that correlation is a weighted average in which each correlation is weighted by the number of people in the study (the sample size, N). The corresponding variance computed across studies is not the usual sample variance, but a sample-size-weighted average squared error

$$S_r^2 = \frac{\sum \left[N_i \left(r_i - \bar{r} \right)^2 \right]}{\sum N_i}.$$

Again the reliability meta-analytic method proposed by Vacha-Haase (1998) departs from the Hunter and Schmidt (1990) method. Vacha-Haase (1998) does not weight the reported reliabilities by sample size, but instead includes sample size as one of the variables in a regression analysis. Like the mean, the Vacha-Haase variance is computed using unit weights.

Step 3: Correct the variance by subtracting the amount attributed to sampling error.

Hunter and Schmidt (1990) outlined steps to estimate the amount of variance due to sampling error and then addressed how to subtract variance attributed to sampling error from the overall variance. Vacha-Haase (1998) did not address any method for partialing-out sampling error from the overall variance.

The Hunter and Schmidt method is a type of random-effects method of meta-analysis. Random-effects methods estimate the variance expected to be observed if the studies were all computed on samples of infinite size. The variance of the distribution of infinite-sample studies is called the random-effects variance component (REVC). In the Hunter and Schmidt method, the REVC is denoted σ_ρ^2. The square root of this quantity is the standard deviation of infinite-sample effect sizes, denoted σ_ρ. The Vacha-Haase method is a fixed-effects method that is closely related to Rosenthal's (1987) method of meta-analysis. In the fixed-effects methods, variability in the infinite sample effect sizes is not estimated. Rather, it is assumed to be zero after accounting for moderators.

Step 4: Correct the mean and variance for study artifacts other than sampling error.

Hunter and Schmidt (1990) corrected the mean and the variance of the study for artifacts that included reliability. The reasoning behind the Hunter and Schmidt method was to cancel-out what they considered to be distracters from the true validity estimate. They used equations based on psychometrics to estimate the correlation between true scores. For test validation (validity generalization), Hunter and Schmidt (1990) advocate correcting for criterion unreliability and direct range restriction in the predictor to estimate a disattenuated mean correlation ($\hat{\bar{\rho}}_{XY}$).

Vacha-Haase (1998) did not address corrections for artifacts; instead she moved on to a moderator analysis. Vacha-Haase's moderator analysis will be discussed in further detail after Step 5 of the Hunter and Schmidt (1990) method is covered.

Obviously Hunter and Schmidt's artifactual correction for reliability cannot be used when meta-analyzing reliability. Perhaps what is less obvious is whether reliability has its own artifacts, and whether reliability artifacts should be uniquely considered and addressed when computing a meta-analysis.

Step 5: Compare the remaining standard deviation to the mean to assess the size of the potential variation in results across studies.

In this step of their meta-analytic method, Hunter and Schmidt (1990) examine the ‘generalizability’ of the results by computing what they called the lower bound of the credibility interval. It is computed (approximately) by: $LB = \hat{\bar{\rho}}_{XY} - 1.96\hat{\sigma}_{\rho}$. The lower bound indicates a threshold below which it is expected that infinite-sample correlations will rarely be found.

Because this step depends upon the random-effects variance component, it is irrelevant to a fixed-effects method such as that used by Vacha-Haase (1998). Therefore, there is no step in Vacha-Haase that corresponds this step five.

Step 6: The Moderator Analysis

Unless the estimate of σ_{ρ}^{2}, once sampling error is subtracted, is sufficiently large, Hunter and Schmidt (1990) advocate abandoning the search for moderators. They note that not all artifactual sources of error can be corrected (e.g., typographical and computational errors), so that $\hat{\sigma}_{\rho}^{2}$ may be positive even though there is only a single true (infinite sample) value of $\bar{\rho}$. If $\hat{\sigma}_{\rho}^{2}$ is sufficiently large, however, tests for moderators may begin. The moderator analysis proposed by Hunter and Schmidt (1990) is to split the data into categories based on the levels of the moderator variable, and then to meta-analyze each category separately. Hunter and Schmidt (1990) do not recommend dealing with the issue of analyzing continuous independent moderator any differently then with dichotomous or multi-level moderators. They point out the using multiple regression to analyze for moderator variables includes too many issues with low statistical power and capitalization on chance, and thus don’t recommend using it (Hunter and Schmidt 1990, pg. 408)

A philosophical difference. Both the Hunter and Schmidt (1990) and Vacha-Haase (1998) methods attempt to account for the observed variance in effect sizes across studies. The two methods look to explain the observed variance in very different ways, however. The Hunter and Schmidt method involves a great deal of attention to artifactual corrections that they expect to explain any differences among observed validity estimates. In other words, the Schmidt and Hunter theoretical position is that the observed variance in validity effect sizes is due entirely to artifacts.

The Vacha-Haase theoretical position, however, is that the variance in observed reliability effect sizes is due to substantive reasons. The main point of the analysis according to Vacha-Haase (1998) is to discover and name those things that cause reliability to differ across situations. While this may not omit the Hunter and Schmidt method from consideration, it does give weight to the thought that other methods may prove to be more suited to the meta-analysis of reliability.

Vacha-Haase (1998) and the studies that followed (Caruso et. al., 2001; Viswesvaran and Ones, 2000; Yin and Fan, 2000) started with the expectation that reliability would vary due to factors other than sampling error. In fact, two of their major goals for meta-analyzing reliability were to "(c) look at the amount of variability in reliability coefficients for given measures and (d) identify some of the sources of variability".

In the analysis that Vacha-Haase (1998) developed, features of the studies that were suspected to add to the variability of the reliability estimate (i.e., moderators) were dummy-coded (i.e. type of reliability coefficient, gender, long vs. short form of the test, language) and then an ordinary least squares (OLS) regression analysis was conducted to explore how the study features predicted variations in the reliability coefficients. Vacha-Haase did not directly address any issues around artifacts and reliability generalization.

Which particular method may be the best approach to reliability meta-analysis becomes even cloudier when the subject of normal versus non-normal distributions is introduced. The Vacha-Haase (1998) and the Hunter and Schmidt (1990) methods shared the assumption that the underlying distribution of the effect size mean estimate is normal or close enough to normal that ‘normalizing’ the data is not necessary. There are those who disagreed.

Fisher's r *to* z*: Should it Be Part of the Meta-Analytic Method?*

Reliabilities are represented as correlations of one test across two times in test-retest methodology. The theoretical sampling distribution of observed correlational values is non-normal in any sample where N is not larger than 500 (James, Demaree and Mulaik, 1986, pg. 446). The distribution is negatively skewed for a positive population mean (rho) and the degree of skew, as well as the kurtosis, increases as the value of rho increases (Fisher, 1954). When rho becomes especially large, as is the case in reliability where rho tends to fall between .60 and .90 (Hogan et. al, 2000), the distribution will remain non-normal even in samples over 500 (James et. al, 1986). Figure one is a graph which depicts the skew in the observed distribution of a large set of reliability estimates based on Hogan et. al, (200).

The sampling distribution of *r*'s is not the only skewed distribution, for example, when rho is considered to be a random variable (as it is in the random-effects case), then the underlying distribution of rho may also reach a ceiling at 1 and thus become truncated and partially skewed. For both reasons, the observed distributions tend to be skewed, probably much more so than validities, which tend to accumulate in the range of .2 to .5 (Brannick & Hall, 2000).

Sawiloswsky (2000), in fact, mentioned these issues as part of his criticisms of the Vacha-Haase (1998) analysis. He noted that, just as previously explained, a reliability coefficient is a correlation coefficient, and as such, may mean a non-normal distribution. He suggested that the Fisher's *r* to *z* transformation should be applied prior to the meta-analysis to ensure a normal distribution. Others have agreed with this observation.

Silver and Dunlop (1987) concurred, when they explained that with the exception of Cronbach's Alpha, reliability coefficients are reported as correlations (*i.e.* the relationship between test and retest, test and similar test, or split-halves of the same test). They further explained that correlations have some difficult statistical properties that may be better handled by using the Fisher's *r* to *z* transformation.

The above examples show that there is currently a debate in the literature as to the correct use of the Fisher's *r* to *z* transformation within the meta-analytic models (Erez, Bloom and Wells, 1996; Hunter and Schmidt, 1990; Silver and Dunlop, 1987). However, Vacha-Haase (1998) clearly did not use this transformation. So again, the question arises: which meta-analytic method is the most appropriate for reliability analysis?

A brief outline of the Fisher's r to z argument is thus outlined next. On the pro-transformation side with Silver and Dunlop (1987) and Sawiloswsky (2000) are Hedges and Olkin (1985) who argued for using the transformation because product-moment correlation coefficients have some undesirable statistical properties, such as a problematic standard error formulation, and an often times skewed distribution. The application of the Fisher's *r* to *z* transformation helps to alleviate those problems by normalizing the distribution and providing for an easier standard error statistic.

On the anti-transformation side of the argument, Thompson and Vacha-Haase (2000) rebutted Sawiloswsky (2000) by explaining that a reliability coefficient is really a "population (or domain) variance-accounted-for statistic" (p. 186), which is estimated by computing the unsquared correlation between scores on observed parallel tests or on a single-test administered twice. They further suggested that because reliability is computed with unsquared *r*-values, the resultant reliability coefficient is also a variance-accounted-for statistic and thus reliability coefficients are usable, as they are, in averaging across studies. Thompson and Vacha-Haase (2000) however, did make a small concession at the end of their explanation, saying that it would be reasonable to the take the square root of the reliability coefficients and apply Fisher's *r* to *z* transformation.

Hunter and Schmidt (2000) also argued against using Fisher's *r* to *z* transformation. They asserted that the Fisher's *r* to *z* transformation produces an estimate of the mean correlation that is upwardly biased and less accurate than an analysis using untransformed correlations. They concluded that the transformation gives larger weights to large correlations than to small ones, resulting in the positive bias. They pointed-out that Fisher's purpose was to create a transformation of the correlation for which the standard error (and subsequent confidence intervals) would depend solely on sample size and not on the size of the parameter.

Silver and Dunlap (1987) refuted Hunter and Schmidt's (1990) position with a Monte Carlo study using the Fisher's *r* to *z* transformation when averaging correlation coefficients. Their results indicated that regardless of sample size, backtransformed averaged *z* was always less biased than a non-transformed *r*. They recommended the use of the *z* transformation when averaging correlation coefficients and particularly when there is a small sample size.

Hall and Brannick (2002) compared the Hedges and Vevea (1998) random-effects model and the Schmidt and Hunter (1990) model, specifically looking at the impact of Fisher's *r* to *z* transformation, in the context of validity meta-analysis. They used a Monte Carlo method to check both the Schmidt and Hunter and Hedges and Vevea credibility intervals against the population credibility intervals. They found that there was a slight difference in means, and some more noticeable differences in credibility intervals. The difference in credibility intervals generally favored the Schmidt and Hunter method. Although the *r* to *z* transformation was not the only difference between Hedges and Vevea and Schmidt and Hunter methods in their analysis, it did contribute to those differences. Brannick and Hall (2000) estimated that if the validity estimates they were analyzing had been even more congregated on the upper-end of the distribution, as they would be in reliability distributions, the differences between the Schmidt and Hunter and the Hedges and Vevea model results might have been even larger.

What remains unclear is if the skewed distribution will create more error in the Vacha-Haase and the Hunter and Schmidt methods where the Fisher's *r* to *z* is not used. It seems likely that it will create more error if the transformation is not used, but this has not been examined empirically yet.

Lipsey and Wilson (2001) developed a random-effects meta-analytic method, which incorporated the Fisher's z transformation. It is possible that the results from this type of approach would be different from either the Vacha-Haase (1998) or the Hunter and Schmidt (1990) methods. How the results would differ and to what extent they would differ needs further investigation.

The Lipsey and Wilson (2001) method also contributes some additional unique analysis of the between-study variance that neither the Vacha-Haase nor the Hunter and Schmidt methods evaluate (Erez et al., 1996; Hedges and Vevea, 1998; Lipsey and Wilson, 2001). Hedges and Vevea (1998) made the argument for methods that incorporate estimates of the random-effects variance components (REVC's). They stated that the modeling of random effects type variability, when that variability exists, would produce a more accurate estimate of the average effect size and the credibility of the interval around the effect-size statistic.

Given the evidence, it is possible that a meta-analytic procedure such as the one used by Lipsey and Wilson (2001), which incorporates the Fisher's *z* transformation, may enhance reliability analysis. The random-effects method, as described by Lipsey and Wilson (2001), has not yet been applied to reliability meta-analysis in any published studies; therefore, the impact of its use remains unknown and worthy of investigation.

Lipsey and Wilson Method of Meta-Analysis

Lipsey and Wilson (2001) employed six basic steps in their meta-analytic method.

1. **Assemble statistically independent effect sizes.** In reliability meta-analysis, all effect sizes are represented as correlations.

2. **Transform *r* to *z*.** Because reliability is represented as a correlation, there are difficulties with the statistical computations; this is especially true of the standard error formula (Rosenthal, 1994). Lipsey and

Wilson recommend applying the Fisher’s *r* to *z* transformation to help correct these issues.

3. **Compute appropriate weights for that effect size.** In the case of reliability meta-analysis the *inverse variance weight* would be applied. Neither the Vacha-Haase (1998) nor the Hunter and Schmidt (1990) method applied this weighting. Lipsey and Wilson argue that because different sample sizes are being compared in a meta-analysis, large sample sizes more closely approximate true population characteristics. Thus, it seems reasonable to weight those sample sizes more heavily in the meta-analysis. A straightforward approach to this would be to just weight by the sample size, as in the Hunter and Schmidt (1990) method. However, Hedges and Olkin (1985) have demonstrated that optimal weights are based on the standard error of the effect size (the standard deviation of the sampling distribution). Because larger standard error equates to a less precise value, the inverse of the squared standard error values are used as the weights. This is called the *inverse variance weight*. For the *z* distribution, the inverse variance weight is (N-3).

4. **Estimate the mean and random-effects variance component.**

5. **Assess the adequacy of mean effect size for representing the entire distribution of effects.** Homogeneity testing is done at this time.

6. **If homogeneity is rejected; then the analyst must choose between three models.** The **Random Effects Model** would calculate the

REVC (V_θ) and then incorporate it into the inverse variance weights and recalculate the mean. However, if the analyst believes that there may be error also due to moderators, then either the fixed effects model or the mixed effects model should be considered. In the **Fixed Effects Model**, similar to the Vacha-Haase analysis, a weighted regression analysis is done to identify significant moderators. The idea is that the moderators will account for all of the variance in V_θ. If however, there is good reason to believe that moderators may only account for a proportion of the random variance and that there may well be a random effects component left after all moderators are accounted for, than the **Mixed Effects** approach is the most appropriate. In the mixed effects model, the REVC (V_θ) is derived and incorporated into the recalculation of the weighted mean. However, as opposed to the pure random effects model, the presence of the moderator requires matrix algebra to estimate the random error variance term. This can be calculated using a SAS macro devised by Lipsey and Wilson (2001).

7. **For Random Effects and Mixed Effects Models, moderators are examined using weighted least squares regression with the corrected inverse variance weights**.

Lipsey and Wilson include both the inverse variance weighting procedure and the Fisher's *r* to *z* transformation in their models. This sets their approach apart from both the Vacha-Haase and the Hunter and Schmidt methods, which were previously described.

What remains unclear is exactly how each of these methods differs in estimating parameters of reliability within the meta-analytic model. A brief overview of each method is presented in table 1.

Finally, one should note that effect sizes are usually reported as a range, or interval, along with the mean. Two different intervals have been used in the literature: confidence intervals and credibility intervals (see Whitener, 1990). Confidence intervals represent the bounds within which, with a pre-defined certainty (usually 95%), the true population mean is expected to reside. This suggests that a true value of rho exists and that the variance observed is due to sampling error.

Credibility intervals, on the other hand, are expected to contain a specified percentage of the distribution of rho, when rho is considered to be a random variable. Credibility intervals therefore represent the range with which rho would fall even if sampling error were not present. Credibility intervals imply that there is not one true population rho, but a range of values differing according to context. Computationally, this difference is represented in the error term used to calculate the interval. Confidence intervals are calculated using the standard error of the mean, usually the square root of variance divided by the square root of the total sample size (or formulations designed to approximate this term). Credibility intervals are calculated using the square root of corrected variance (after sampling error is accounted for) without a denominator. Credibility intervals are usually larger than confidence intervals, and can be calculated only when a random- or mixed-effects model is assumed.

A Closer Look at Each Method Using a Small Data Set

The following example is based on numbers that are fictional but plausible in reliability literature.

The *N*'s represent the number of participants per study. The r_i's represent the reported test-retest reliability for each study. In addition, there is included for each study, a time interval between test and retest, derived using a logarithmic function that simulates the decay of reliability over time.

This example is intended to provide the reader with a better understanding of the computations and expected differences between methods. Although the original methods were presented in the steps given by each author, the following examples will share a similar format to provide for better comparison between methods.

We will now illustrate the three main meta-analytical methods described in the preceding section with a set of test-retest reliability data for the Mason-Brannick Non-Existent Personality Test. The data are fictional and designed to illustrate the techniques and in general, we do not always expect to see an association between N and the size of r. Table 2 shows the sample test data.

Vacha -Haase method. For step one, each method computes effect size statistics (i.e. reliabilities) and finds the average effect-size across studies.

In Vacha-Haase this mean-effect-size is computed as a unit-weighted average by the formula: $\bar{r} = \frac{\sum r_i}{K}$. Using the sample data, Vacha-Haase calculates $\bar{r}$ = 0.76, which is a straightforward calculated mean. In Vacha-Haase, the next step is to construct a box and whisker plot to represent the distribution of effect sizes which is shown in figure 2.

Variance is then computed with the standard variance formulation, the mean squared deviations from the mean, $\hat{\sigma}^2 = \frac{\sum \left(r_i - \bar{r}\right)^2}{n-1}$ = 0.05.

Although Vacha-Haase didn’t report confidence intervals, they have been computed here for the sake of comparison with the other methods. The confidence intervals are shown below calculated as $\overset{\omega}{r} \pm 1.96(SEM)$ where SEM is the previously mentioned standard error of the mean, calculated as $\frac{\hat{\sigma}}{\sqrt{n}}$.

Hence the interval is 0.7633 ± $\frac{0.2160}{\sqrt{6}}$

Confidence Intervals for Vacha-Haase method using sample data are outlined in table 3.

The final step in the Vacha-Haase method is to compute an ordinary least squares regression to check for moderator effects. In this case, the **unweighted** r_i is regressed on interval in days, and N (number of subjects per study). The results of this regression are displayed in table 4.

The analysis would indicate that the interval between test-retest is a significant moderator as shown by the *t* of –5.8, but that sample size is not, because that *t* was not significant. Here the Vacha-Haase method would end (although there are well-known problems with regression analysis such as collinearity, their discussion is omitted from the illustration for brevity and clarity). The analysis would show that time-interval moderates the value of the test-retest reliability coefficient. Because the analysis is fixed-effects, there is no estimate of the random-effects variance component (REVC) or any additional variance that is not accounted for by the moderator(s) in the analysis.

The Hunter and Schmidt method. Table 5 represents the data with the necessary calculations for the Hunter and Schmidt method. The *r*, *N* and Time-Intervals are the same as the previous Vacha-Haase example.

The weighted mean in the Hunter and Schmidt is equivalent to $\bar{r}$, and is calculated as $\bar{r} = \frac{\Sigma(N_i r_i)}{\Sigma N_i}$, read as average reliability weighted by N. For this sample data, $\bar{r}$ is equal to $305 \div 372 = 0.819892$ or .82. This mean is then used to calculate the observed variance.

Observed variance is calculated using the formula $s_r^2 = \frac{\Sigma\left[N_i(r_i - \bar{r})^2\right]}{\Sigma N_i}$, which for this example is equal to $10.0468 \div 372 = 0.027008$ or .03.

The next step for Hunter and Schmidt is to estimate sampling-error variance. The formula, as previously stated, is in the form: $\sigma_{\mathrm{e}}^2 = \left(1 - \bar{r}^2\right)^2 / \left(\bar{N} - 1\right)$. Substituting the previously obtained weighted average gives a value for sampling-error of 0.001761.

Estimated variance around the population mean (rho) is then computed by subtracting the estimated sampling-error from the observed variance, $\hat{\sigma}_\rho^2 = s_r^2 - s_e^2 =$ 0.027008-0.001761= 0.025247 or .03. This number is the estimate of the random-effects variance component. (Note that Hunter and Schmidt use the symbol σ_ρ^2 to refer to the random-effects variance component (REVC), but Lipsey and Wilson refer to the same quantity as V_θ.) Hence the standard deviation is $\hat{\sigma}_\rho = \sqrt{.025247} = 0.158893$ or .16.

Credibility intervals are now constructed using the weighted mean and $\hat{\sigma}_\rho$ with the appropriate z-value (1.96 for 95% confidence interval) using $\bar{r}$ ± 1.96 ($\hat{\sigma}_\rho$) = 0.819812 ± 1.96*(0.158893). This represents a credibility interval since it is the expected range of theoretical values, not the interval expected to contain the *mean* and it is calculated after sampling error has been accounted for. Table 6 shows the credibility interval for the S-H results.

This illustrates that there is sometimes a problem with the estimate of the upper limit of the distribution with the Hunter and Schmidt method. The maximum admissible or theoretically possible value of the correlation is 1.0. The best upper estimate in such a case is arguably 1.0 rather than 1.13. Such a result also suggests that the normal distribution may not be the best approximation for reliability distributions.

To approximate the confidence intervals for the Hunter and Schmidt method, the standard deviation could be divided by the square root of *k* (the number of studies). Because this is in the random effects scenario, the resulting confidence interval is expected to contain the mean of the random variable rather than the single value of the population mean. In symbols, we expect the confidence interval shown below to contain $\overline{\rho}$ rather than ρ. In the random-effects case, standard error of the mean would be $\frac{0.159}{\sqrt{k}}$ or $\frac{0.159}{\sqrt{6}} = 0.065$. The confidence intervals are computed as $\bar{r}$ ± 1.96 (0.065). Table 7 shows these approximate confidence intervals.

Because this data set has only one moderator (interval between test and retest), Hunter and Schmidt would probably separate the studies based on the level of the moderator, such as over 1 month, 2 weeks and less then 2 weeks (high, medium, low).

Then each set of studies would be meta-analyzed independently. The Hunter and Schmidt process would continue to divide studies into categories based on moderators until there was no (or small) remaining variance left unaccounted for.

The Lipsey and Wilson method. Table 8 contains the same sample data and calculations as before. However, the first step in the Lipsey and Wilson method is to transform the study effect sizes using Fisher's *r* to *z*. The transformation results are in the column labeled *z,* derived for each *r* using the transformation formula:

$$z = .5\ln\left[\frac{(1+r)}{(1-r)}\right] = \text{atanh}(r)\cdot$$

The next step in the Lipsey and Wilson is the same as the first steps in the Vacha-Haase and Hunter and Schmidt methods, which is averaging the effect sizes. Similar to Hunter and Schmidt, Lipsey and Wilson calculate a weighted mean. However, in addition to using the *z*-values, they use the inverse variance weight (N-3), calculated and labeled *w* in the table above. Thus $\bar{r}_z = \frac{\sum_{i=1}^{k} w_i z_i}{\sum_{i=1}^{k} w_i} = \bar{z}$ $= \frac{467.80}{354}$ = 1.32147 or 1.32.

As an example of the standard error of the mean computed in the Lipsey and Wilson methodology, s $= \sqrt{\frac{1}{\sum w}} = \frac{1}{\sqrt{354}}$ = 0.053149. This is interesting because it involves the summation of the inverse variance weights. This might be recalculated depending on the outcome of the Q statistic, to be explained next. If the Q statistic were not significant, the above result would be used to calculate the confidence intervals.

For the random-effects method, Lipsey and Wilson consider the variability of the effect sizes. They both test for the homogeneity of effect sizes in the population and estimate the variance of the infinite-sample effect sizes in the population. The estimate of the variance of infinite-sample effect sizes may or may not be conditional on a significant test of homogeneity of effect sizes, depending on the researcher's choice. The homogeneity test, Q, is used in the calculation of the variance estimate for the infinite-sample effect sizes.

The homogeneity test involves computing Q, which is distributed as chi-square when the null hypothesis is true. The null hypothesis is that all of the population effect sizes are equal, that is, $\rho_1 = \rho_2 = ... = \rho_k$. Q is calculated as a weighted sum of squares, thus: Q = $(\sum w_i z_i^2) - \frac{(\sum w_i z_i)^2}{\sum w_i}$ = (in our example) 681.22 - $\frac{(467.8)^2}{354}$ = 63.04.

If Q exceeds the chi-squared value within the appropriate degrees of freedom (number of studies less one), then the null hypothesis of homogeneity is rejected. If it's rejected, then there is variance over and above sampling-error that may be accounted for by moderators. In our example, there are 6 studies, and therefore 5 degrees of freedom for the Q statistic. The critical value of chi-square (α=.05) with 5 *df* is 11.07, so we can reject the null in our example. The conclusion, there is variance unaccounted for by sampling-error alone.

The analyst now has three models from which to choose to evaluate the variance. These are, as previously described, a pure random effects model, a pure fixed effects model and a mixed model. The fixed effects model would assume that the unaccounted for variance in *r* is due to systematic variables, (i.e., moderators). In this model there is no random error term computed, since it is assumed to be zero. Therefore, similar to the Vacha-Haase method, a regression is run. However, in this case it is weighted by the inverse variance weight and is performed by regressing the weighted, transformed *z*-values on the postulated moderator. Of note is that fixed effects models are less favored in the current literature due to the high type I error rates, if in fact there is a random variance component (Lipsey and Wilson, 2001). Therefore, the focus in this paper will be on the remaining two models, random effects and mixed, which account for the random error variance component (REVC).

In both of these models, a calculation is made for the REVC, now denoted as v_θ. This random error variance term is then added to the initial observed variances, new inverse variance weights computed, and, finally the weighted mean is recomputed using the new inverse weights.

The calculation of the random-effects variance, denoted v_θ, is as follows in the pure random effects model:

$$v_\theta = \frac{Q-(k-1)}{\sum w_i - (\sum w_i^2 / \sum w_i)}$$, which in our example, means that

$$v_\theta = \frac{64.03-(6-1)}{354-(23188/354)} = 0.201187$$

The rounded value (.20) is the random-effects variance component for the Lipsey-Wilson method. This value is analogous to the value of .03 obtained using the Hunter-Schmidt method. Note, however, that the two numbers are not directly comparable. The Hunter-Schmidt estimate is a variance of a distribution in *r*, but the Lipsey-Wilson estimate is a variance of a distribution in *z*. There is no simple transformation of the variance in *z* that will make it directly comparable to the estimate in *r*.

In the random-effects case, variance (uncertainty) comes from two different sources, (a) finite sample size from individual studies, that is, sampling error, and (b) variability in the true or infinite-sample effect sizes. Proper weighting of studies to best estimate the mean in such cases must consider both sources. Therefore, the inverse variance weight, which was previously calculated as n_i-3, is now recalculated with v_θ added to the variance term. Thus, $v_i^* = v_\theta + v_i$. As an example, for the first study, number 1, v_1 initially was 1/(n-3)= 1/82= 0.012195. The new variance, v_1^*, becomes $v_1 + v_\theta$ = 0.012195 + 0.201187 = 0.213382. Thus, the new inverse variance weight will be 1/(v_1^*), or 1/0.213382= 4.6864. New (revised) weights are calculated for each study. The revised inverse variance weights are then used to calculate a revised meta-analytic weighted effect size mean.

When using the pure random effects model, all of the unaccounted for variance is assumed to be random. Thus, all of the observed variance other than sampling error is incorporated in the v_θ computation. This assumption is problematic when moderators are present and unaccounted for. The question becomes how the analyst tests for moderators and still allows for a reasonable random error component.

The mixed effects model allows for both moderators and remaining random-effects variance. In the mixed effects model, the analyst assumes that there is some variance in *r'*s due to moderators and some due to a random error component (over and above sampling error). In the mixed effects model the computation of v_θ is based on complicated matrix algebra formulations since the estimate is based on residual variability rather than total variability. The explanation of the matrix procedures used is beyond the scope of the present study. However, macros have been developed in both SPSS and SAS to handle the matrix calculation and the recalculated mean effect size (Lipsey and Wilson, 2001).

Finally, as opposed to the pure random effects model, but similar to the fixed effects model, the final step would involve a weighted regression analysis using the new inverse variance weights. The output for the data presented in the table above using the mixed effects model macro for SPSS and a method of moments estimate for v_θ will be outlined next. In the mixed-effects model, a random-effects variance component (REVC) is computed after taking the moderator into account. For the current data, the estimate is v_θ = .0294. Using this v_θ to recalculate the inverse variance weights will result in a new mean with confidence intervals between 1.07 and 1.41, as shown in table 9.

Of course these numbers in those confidence intervals are still in Fisher's *z* and need to be back transformed into *r* to make them comparable with the previous methods' results. Table 10 shows what the confidence intervals would be once they are backtransformed.

A regression analysis using the inverse variance weighted *z*'s, known as a weighted least squares regression is also run in the mixed model. In this case the *z*'s would be regressed on "time interval between tests" variable, weighted by the inverse variance weights (w_i).

Using the sample data, the SPSS weighted least squares regression output is presented in table 11.

These results indicate that the moderator, “interval of time between tests”, is a significant contributor to the variance.

Lipsey and Wilson do not provide an exact formula for calculating credibility intervals. Standard deviation of the population, which is used to construct credibility intervals, can however be approximated by multiplying the revised standard error of the mean term by the square root of *k* (number of studies). This looks like $0.089 * \sqrt{6} = .218$. This is actually only a rough approximation for these results because with the continuous moderator influencing the variables, credibility intervals can be constructed around any point that falls on the regression line. However, this is an approximation of the average point on that line and the credibility around it. In reality credibility probably wouldn’t be calculated at all, but for purposes of comparison, we will use this estimate. A credibility interval can now be calculated for the range of *z* scores, back transformed into *r* scores as displayed in table 12.

Comparison of methods. The results from the Vacha-Haase, Hunter and Schmidt and the Lipsey and Wilson mixed effects model are presented in the table 13 to provide for an easy comparison of the results across methods. This table shows the confidence intervals for each method.

In table 14 the credibility intervals for the Hunter and Schmidt and the Lipsey and Wilson mixed effects outcomes are displayed.

Even with this limited data, it becomes clear that there are differences in the estimated population parameters between the methods. Which one is most correct is difficult to determine however, because the true population values are unknown.

What model should be used for Reliability Meta-Analysis? The three models of meta-analysis summarized under common steps in the above tables share some common features, but also contain unique features. Vacha-Haase's method seemed to most closely resemble that of Rosenthal (1987) in his explanation of how to combine correlations and compute resulting variance in a fixed-effects model. However, even Rosenthal incorporates the Fisher's *r* to *z* transformation as a necessary part of the method, making his method an imperfect match as well.

Perhaps what is most important is not what Vacha-Haase (1998) and others *have done* so far, but the improvement of the methodology around the concept of meta-analyzing reliabilities of tests and measures for future research. This study is an attempt to examine the existing methods of meta-analysis of reliability estimates with an eye to informing future methodological choices.

The goal of this research was to determine which method is the better statistical approach for the meta-analysis of reliability data. The study compared the Vacha-Haase (1998) method, the Hunter-Schmidt (1990) method, and the mixed effects method as outlined in Lipsey and Wilson (2001) against one another and against a known standard to inform researchers. This portion of the study also included an analysis of the impact of the Fisher's *r* to *z* transformation on reliability coefficient analysis in hopes of answering the question of whether the transformation is helpful in reliability meta-analysis.

The study also examined the influence of the choice of weights, whether sample size (as in the Hunter and Schmidt example), inverse variance weights (as in the Lipsey and Wilson method) or whether sample size should be treated just like any other moderator influence (as in the Vacha-Haase model). Finally, the study compared weighted and unweighted regression procedures to examine impact of the choice of procedures on the probable outcome of the meta-analysis.

Method

In a "real world" meta-analysis there is no way to know which estimates of the population characteristics are closest to their true values. In an effort to distinguish the best method for meta-analyzing reliabilities, a Monte Carlo simulation was used. The advantage to using a Monte Carlo simulation is that it provides a way to set the population characteristics *a priori* and then to compare each method's outcomes to the population values.

The Monte Carlo simulation was used to compare the results of the different meta-analytic methods when the samples are drawn from the typically skewed reliability sampling distribution.

The nature of the reliability distribution, especially as it becomes truncated and skewed at the upper limits, and its impact on the estimates of the population characteristics is at the heart of the *r* to *z* transformation debate. The results from the Monte Carlo simulation shed some light on whether transforming the reliability coefficients to the more normal *z* distribution, provides for better estimates of the parameters (mean and variance of infinite-sample reliabilities).

The results show how well each of the three approaches recovers known means and variances under several realistic conditions.

Also included in the simulation was a moderator variable that functions similarly to the moderator of time between test and retest intervals in the previous example. This provided insight into the relative merits of Vacha-Haase and the Lipsey and Wilson methods when a continuous moderator variable is present.

The bias and standard error of slope estimates for each model were examined, as well as Type I and Type II error rates for slope estimates. The point of these analyses was to highlight the advantages and disadvantages of each method and to make recommendations on when each approach is most appropriate in meta-analyzing reliabilities.

Monte Carlo Study

Study overview. This study incorporated a Monte Carlo simulation where mean reliability ($\bar{\rho}$) and variability (σ_{ρ}) of infinite-sample studies were manipulated. The number of studies in each meta-analysis (*k*) varied systematically and the sample size per study (*N*) was generated as a random variable. Data (simulated studies) were generated under each condition. Simulated studies were then meta-analyzed. Data generation is described in two parts, one for fixed-effects and one for random-effects. Data analysis is also described in two parts. Part one of the analysis examined estimates of the mean and variance of infinite sample effect sizes provided by the three different methods of meta-analysis (Hunter-Schmidt, Lipsey-Wilson and Vacha-Haase). Part two of the analysis examined moderator effects using two of three methods (Vacha-Haase and Lipsey-Wilson).

The choice of parameters. In part one, the three methods were compared against one another for their estimates of the mean reliability and variability of a known set of “true population” values. The advantage of the Monte Carlo study is that a researcher can chose what the population parameters are. In this study, the data emulated real-world conditions as much as possible to provide for a useful comparison of the meta-analytic methods. Thus, values were chosen for the population mean and variance that were based on a real-world example. The values chosen were based on a cognitive ability test known as K-TEA/NU. Based on the information from the test-retest data from the K-TEA/NU, a moderator was also uncovered. A short discussion about this moderator is necessary to describe how the population values were chosen and how they relate to these real-world circumstances.

Decay of Reliability Over Time

Time between test-retest, measured in days, is known to have a moderating effect on test-retest reliability. Typically, as time between tests increases, the reliability estimate decreases because participants change more as time increases (Viswesvaran et. al., 1996). Also, if test-retest rather than alternate forms data are collected, participants tend to remember their responses to specific items in the earlier administration. Such memories can inflate the reliability estimate, particularly for short retest intervals, causing the reliability to appear much lower over longer time periods (Nunnally & Bernstein, 1994).

Reasonable values used in the simulation were based on the test-retest data associated with different time intervals from the cognitive ability test K-TEA/NU (AGS Publishing, 2002). The K-TEA/NU test data gave a range of .97 - .80 over an interval of 3- 35 days. The assumption was made that the decay in reliability, like most time-dependent decay functions, is represented well by a logarithmic function. Thus, reliability is linearly related to log of time with a negative slope. The chosen form of the function was:

observed reliability = (maximum reliability) -.04*ln*(time in days), (1)

where *ln* is the natural logarithm. Figure 3 illustrates the function. The upper line corresponds approximately to the data for the K-TEA/NU, for the function $r_{xx} = .95-.04ln(t)$, where *t* ranges from 1 to 35 days.

In order to come up with the value to use as the mean of the population, the mean and variance of the function was calculated. The mean reliability for this function is .84, and the standard deviation of reliability is .03.

The second line was introduced to increase the generalizability of the findings to measures such as job satisfaction that are somewhat less reliable than professionally developed cognitive ability tests and thus would have a lower mean population value. The bottom line in Figure 3 is an example of what might be seen in a job satisfaction measure. This line starts at .85 rather than at .95; its mean reliability is .74 and its standard deviation is also .03. Figure 3 shows what the decay of reliability over time looks like graphically for the two different estimates of reliability.

As previously mentioned, this moderator was derived from actual data. It appeared to be a reasonable choice for this study because the length of time between test and retest is almost certain to influence the magnitude of the reliability estimate and because time is a continuous variable. This is important because the Hunter and Schmidt method of breaking moderators down into discrete groupings is obviously much more difficult in such a scenario. Because continuous moderators are likely to appear when analyzing reliability, they deserve close consideration.

Part One: Data Generation for Fixed Effects

Data were generated based on the means of the two lines in Figure 3 (.84 and .74). In this fixed effects case, the only source of variance was sampling error. The two values of that ρ were .84 and .74, which again are equal to the two means in the conditions based on the K-TEA/NU data and in which reliability decays over time.

Number of studies (k). The number of studies (*k*) included in each meta-analysis was set to values of 10, 50 and 100. These values were selected to show what happens to the analysis as the number of studies increases. Meta-analyses of large numbers of studies are rare, so 100 appeared to be a reasonable maximum. Because reliabilities are often under-reported in the literature (Vacha-Haase et al, 2002; Yin and Fan, 2000; Whittington, 1998), it is possible to have reliability meta-analyses that are conducted on a small number of studies. This maybe especially true if moderator analyses are conducted according to the Hunter and Schmidt (2000) method, where the studies are divided according to the moderators and then each new grouping is meta-analyzed. Thus, a meta-analysis sample size of 10 studies is also reasonable.

Sample Size (N). The sample size (N_i is the sample size of each study) is directly related to the magnitude of sampling error. Hunter and Schmidt and Lipsey and Wilson both assume that studies with smaller sample sizes are associated with larger (sampling-error inclusive) variance terms. Thus, both methods incorporate a weighted mean as an estimate of the parameter. The weights are proportional to sample size, so that studies with larger sample sizes are given more weight. Vacha-Haase makes no a priori assumptions about sample size and instead incorporates sample size as another variable in the moderator analysis. Therefore, sampling error plays a very different role in the method used by Vacha-Haase than by the other two approaches. Following Hall and Brannick (2002), the sample size per study was drawn from a normal distribution with a mean of 125 and a standard deviation of 25, subject to the restriction that samples meet a minimum of 50. Such a scheme allows samples to vary, but still be large enough to estimate correlations with some accuracy. Such sample sizes are thought to mirror samples taken in current testing programs.

Number of repetitions. Steele et al. (2002) pointed out that some Monte Carlo research uses 10,000 to 25,000 repetitions. However, at that magnitude, millions of separate data points are generated. It was unlikely that this comparison of methods needed quite that many data points to provide clear data on which method most closely approximates the population parameters. Thus, this study incorporated 1,000 repetitions, that is, 1,000 simulated meta-analyses per condition.

Overview of data generation. The data for a single study in Part One were generated in the following manner. In the fixed condition, the value was either .74 or .84. In the fixed condition, there was no variability of infinite-sample effect sizes. Then a sample of size N_i was drawn from the infinite-sample reliability, resulting in an observed study to be included in a meta-analysis. Data were generated using this process for subsequent studies until the required *k* studies (10, 50, or 100) had been generated. Once the required *k* studies were generated, then they were meta-analyzed by each of the three methods. For each condition (value of rho and *k*), 1000 replications were simulated and meta-analyzed.

Part Two: Data Generation in Random-Effects and Mixed-Effects Conditions

One of the major reasons Vacha-Haase first began to meta-analyze reliability estimates was because many researchers were ignoring the possibility of moderators and using the same reliability estimate regardless of the testing situation. Vacha-Haase made the argument that when moderators are present, researchers should consider their impact on their current study. For example, the research should not apply a retest estimate based on a 3-day interval to a situation in which the retest interval is 35 days. Obviously the initial 3-day estimate would be too large. In the very least, some comment should be made regarding the possibility that the test is less reliable over long test-retest periods.

As previously explained, the moderator chosen was assumed to be time decay with a linear relationship between the population rho and ln(t). To make this moderator even more true to real-world data, an additional error component was introduced into the data generation.

The new equation incorporated an error component drawn from a normal distribution with a mean of zero and a standard deviation of .03. Thus the revised moderator equation is:

$$r_{xx} = \text{Maximum} -.04\, ln(\text{t}) +.03\text{e} \qquad (2)$$

The result of adding the error term is to make the decay function somewhat 'fuzzy.' Adding the error term also makes the simulation correspond to the mixed-effects scenario. In a mixed effects scenario, a moderator explains some of the infinite-sample effect size variance, but a part still remains unexplained. This is the scenario that the Lipsey and Wilson method addresses.

Overview of Data Generation

In part two, the data for a single study was generated in the following manner. The time delay between test and retest was sampled from a uniform distribution between 1 and 35 days. The value of time was used to generate infinite sample reliability for that study. In this mixed condition, the value was [either .85 or .95] -.04ln(time)+.03error. Then a sample of size N_i was drawn from the infinite-sample reliability, resulting in an observed study to be included in a meta-analysis. Data were generated using this process for subsequent studies until the required *k* studies (10, 50, or 100) had been generated. Once the required *k* studies were generated, then they were meta-analyzed by all three methods. For each condition (distribution of rho and value of *k*), 1,000 replications were generated and meta-analyzed.

Summary of Data Generation

The data were generated in either a fixed-effects (Part One) or mixed-effects (Part Two) scenario. In both scenarios, the mean value of rho was either .84 or .74. The number of studies was 10, 50 or 100. For each study, N varied essentially randomly. In the fixed-effects scenario, the only source of variability in effect sizes was sampling error. In the mixed effects scenario, the sources of variability included sampling error, a moderator, and an additional random-effects variance component. Table 15 shows a summary of the data generation parameters. For each cell of results (shown in Tables 2 through 6), 1,000 replications were generated. For each replication, all three methods of meta-analysis were used to produce an estimated mean and random effects variance component (all Vacha-Haase random effects variance components are zero).

Analyses

Part One: Mean and Variance. Part One compares the three methods (Hunter-Schmidt, Lipsey-Wilson, and Vacha-Haase) in their estimates of the mean and variance of the infinite-sample effect sizes. For each method (Hunter-Schmidt, Lipsey-Wilson, and Vacha-Haase), the mean and standard deviation of the estimates over the 1,000 trials are reported. For methods that produce unbiased means, the method producing the smallest standard deviation is preferred. For each method, a root mean squared error (RMSE) was also computed by subtracting the parameter (known in the Monte Carlo program) from each estimate, taking the square the result, and then finding the mean and finally taking the square root over the 1,000 trials. In general, methods with smaller RMSE are preferred as a small RMSE indicates that the estimates are generally 'close' to the parameter. RMSE can be used to evaluate the quality of the estimator even if the estimator is biased. The results were summarized in Table 2.

Part Two: Moderator Analysis. In the context of meta-analysis, a moderator variable can be defined as a systematic difference among studies that might explain differences in the strength or direction of relationships between the variables of interest (Steel and Kammeyer-Mueller, 2002). Recently, Steel and Kammeyer-Mueller (2002) compared meta-analytic moderator estimation techniques using a Monte Carlo study. They found that the weighted-least-squares (WLS) multiple regression was the best choice because it is largely unaffected by multicollinearity and heteroscedasticity. Interestingly, they found that the Hunter and Schmidt suggested hierarchical-subgroup-analysis provided the least accurate results among all the methods they analyzed. Because this method fared so poorly and because it does not deal well with continuous moderator variables, a decision was made not to incorporate the Hunter and Schmidt method in the moderator piece of this study's analysis as the results were not likely to provide additional valuable information.

Vacha-Haase (1999) used an ordinary-least-squares (OLS) regression analysis. In her method, the effect sizes are unit weighted. This is modeled after a method suggested by Glass (1977). In the present study only one moderator, time-interval was incorporated. However, Vacha-Haase also included sample size as part of the moderator analysis. Therefore, following Vacha-Haase's example, both sample size and time interval were analyzed as moderators in this study.

Lipsey and Wilson advocated the weighted-least-squares (WLS) multiple regression. Given Steele and Kammeyer-Muller's findings, they seem to have incorporated the most robust methodology for meta-analytic moderator analysis, at least when multiple moderators are present. Additionally, Lipsey and Wilson estimate the impact of both moderator variance and random variance in the mixed effects model. However, WLS regression incorporates sample size in the weights, not as a moderator.

Because of the difference in methods, the Vacha-Haase and Lipsey-Wilson methods differ in both the weights and the set of independent variables. It is therefore of interest to separate the issue of weights from the issue of independent variables. Thus, another analysis of the data was added. In this analysis, unit weighted OLS regression was used without sample size as an independent variable.

Unfortunately for purposes of comparison, the Lipsey-Wilson method uses both WLS regression and the *r* to *z* transformation. Therefore, differences between the methods could be due to weights, the transformation, or both. Further complicating matters is that the regression estimates (slope and intercept) for the Lipsey-Wilson method are in the units of *z*, not *r*. In other words, the Lipsey-Wilson regression estimates apply to transformed data, but the Vacha-Haase estimates apply to the observed data. The two estimates are not directly comparable. To partially disentangle the weights from the transformation, a third method was also added, a weighted regression in which the untransformed values of *r* are weighted by the sample size (N_i). Thus, unit-weighted OLS could be compared to WLS in *r*, and both could be compared to WLS in *z*.

Although the metrics of *r* and *z* prohibit direct comparisons of the variance of the estimators, Type I and Type II error rates for the approaches were directly compared across approaches. For each of the four methods (Vacha-Haase, OLS, WLS*r* and WLS*z*) the slope relating reliability to time delay was computed and tested for significance. The OLS method is known to have an exact Type I error of .05 at alpha = .05, so this provided a check on the accuracy of the program.

Under the conditions in which time delay has an effect (mixed-effects data), the results were used to compute the Type II error rates (or conversely, power) for each of the methods. Methods that actually produce the Type I error rates specified by researchers and also show the maximum statistical power are preferred. Results for both Type I and Type II errors for each method are presented in Tables 21 and 22.

Results of Part Two inform researchers' decisions about the method of analysis for moderators of reliability estimates. Specifically, the results showed the effect of (a) unit weighted OLS versus WLS regression and (b) the effect of the *r* to *z* transformation. Of specific interest to the analysis of reliability data, results also showed the effects of the

Vacha-Haase choice of *N* as a predictor on the error rates for the slope of reliability on time delay.

Results

Overview

In parts 1 and 2 of this study, the three different meta-analytic methods were compared to determine how accurately they would estimate the preset population parameters. Table 15 summarizes the population parameters and can be used as a reference for the remaining tables.

Part One

Mean and variance. In Part One of this study, a Monte Carlo procedure was run and each of the three different meta-analytic techniques was computed. A thousand repetitions for each combination of the two means and the three levels of *k* were calculated, giving a total of six conditions with 1,000 data points in each condition for each of the three methods. As previously explained in the Method Section, the two mean levels were set to approximate the means of the moderator function in Part Two in order to facilitate comparisons. The means were .74 and .84. For each of these two mean levels, the *k* (number of studies) was set to three different levels, namely 10, 50 and 100. Part one is the 'no moderator' or 'fixed' condition, thus the standard deviation is set to .00. This means that the only error incorporated in the local studies was sampling error.

The three different results reported for each condition under each method are: the grand mean effect size statistic over the 1,000 trials, the standard deviation of that mean, and the root-mean-square error. The root-mean-square errors are calculated as the square root of the mean of the 1,000 squared deviations from the population mean (not from the grand mean effect size). Table 16 lists these results.

Each of the three methods produced mean effect size estimates that very closely resembled the parameter. However, the Vacha-Haase and Schmidt and Hunter results consistently underestimated the mean, while the Lipsey and Wilson method consistently overestimated the mean. Such results are consistent with what we know about the sampling distribution of the correlation. Specifically, when ρ is positive as it is in this dissertation, then *r* is a biased (conservative) estimate of ρ, and *z* is a biased (liberal) estimate of the same quantity.

The standard deviations around the means give some indication of how much variance exists in the estimate of the mean across samples. Here again the results are very close across methods. However, the Lipsey and Wilson method does have a consistently lower standard deviation than either of the other two methods by about .001. The standard deviations are larger in the lower *k* conditions and become smaller as *k* increases. Although the results in Table 16 appear to favor the Lipsey-Wilson method, it is difficult to draw any firm conclusions about the use of one method over another, because the differences in the standard deviations are so small. In fact, the methods are seemingly interchangeable in this condition.

The root-mean-square error result often provides additional information that allows a researcher to choose the most appropriate method. The best method would be the one producing the smallest deviation from the population mean as measured by the RMSE. The RMSE’s for these three methods are very close. However in the conditions where *k* is equal to 10, the Lipsey and Wilson method produces a consistently smaller RMSE, suggesting that when there are only a few studies to meta-analyze, the L-W method may be the best approach. This finding is also consistent with what we know about the sampling error of the estimator of the mean. Hedges (1982) has shown that 'inverse variance' weights produce estimates of the mean that are consistent and also have the smallest standard error of any set of weights. As the number of studies increases, all reasonable weighting schemes (including unit weights) tend to produce the same estimate of the mean. With small numbers of studies, however, the choice of weights can be important.

The results for different levels of *k* are also interesting to note. In both mean populations, when *k* is equal to 10, the standard deviations and RMSE's are noticeably higher across all three meta-analytic methods. The larger sampling variance is because of sampling error due to finite *k*; this is what Hunter and Schmidt (1990) called 'second order' sampling error. The smaller the *k*, the less opportunity for discrepant studies to cancel one another out; thus the mean from a small number of studies may not be a very good estimate of the population value. It appears that in meta-analytic research with a small number of studies, researchers need to be much more aware of the potential variance in their results. This point will be continuously supported throughout this study.

Part Two

Mean and variance with the introduction of a moderator. In part two of this study, the same Monte-Carlo procedure was run for each type of meta-analytic method. However, a moderator equation in the form:

$$\rho_i = \rho_{\max} - .04\log_e(t) + e \qquad (2)$$

where ρ_i is the local population value of reliability at test-retest time *t*, $\rho_{\max}$ is the theoretical maximum test-retest reliability in which retest is immediate, *t* is time in days to the retest, and *e* is a normally distributed error term with mean of zero and standard deviation of .03. The equation provides a form for the decay of test-retest reliability as time to retest increases.

As previously discussed in the method section, this moderator equation is a model of a 'real-world' time decay in test-retest reliability estimations and is used to provide a realistic approximation of what happens when moderators impact the magnitude of the effect size (in this case, time delay affects the size of obtained reliability estimate).

The .03 error term in the moderator equation is additional error that is added to the local parameter. This is intended to model random error due to unknown sources or context effects unanalyzed in the meta-analysis. In this case, the amount of the random error is almost exactly of the same magnitude as the standard deviation of the moderator, which has a mean of approximately either .84 or .74 and a standard deviation of .03. Because of the operation of the moderator, the distribution of r_i is only approximately normal (see Figure 5). Even without sampling error (see Figure 6) the distributions are positively skewed. The impact of the two independent sources of variance in ρ_i will be discussed further when looking at the regression results.

In Table 17 the results of each of the three methods are presented exactly as before with the mean, the standard deviation around the mean and the root-mean-square error. However, the pattern in the results is not the same due to the impact of the moderator and error term.

In looking at these results it is first important to remember that when a moderator is present, there is more than one true population. In fact, there are many populations, each with a unique mean value. This is one of the reasons why it is important to discover the possible moderators of reliability. As indicated in the review of the literature by Vacha-Hasse, Ness, Nilsson and Reetz in 1999, many researchers are reporting reliability estimates that are not based on the actual sample in question. When such is the case, and a moderator is present, then the reliability estimate used by the researcher will not correspond properly to the reliability of the data in the local study and the conclusions reached in the local study may be erroneous.

In the case of Table 17, a known moderator is present that would produce a different mean for every possible unit of time for test-retest interval (days between 1-35). In order to present a comparable view of the data, the mean value for the moderator function was computed. The two mean values for each of the moderator conditions are .74 and .84. All of the root-mean-square errors are therefore computed based on these hypothetical mean values.

The r *to* z *transformation.* The evaluation of the meta-analytic results becomes even more difficult when the *r* to *z* transformation is applied, as it is in the Lipsey and Wilson method. This is because the mean of the *z* values backtransformed into *r*'s is not the same mean value as averaging the *r*'s without transformation, because the *r* to *z* transformation is nonlinear and increasingly steep as *r* increases. If ρ_i has a distribution such that the mean and nearly all values in the distribution are positive (as it does in this case), then the distribution of z_i will be positively skewed, particularly if the mean of the distribution is large (as it is in this case). The positive skew will tend to pull the mean of the distribution upward and result in an overestimate when the value is back transformed to *r*. In Figure 5 the same distribution is shown as both ρ_i and z_i to illustrate this point. Note that in Figure 4, the *r* to *z* transformation appears to be working well; the distributions in z appear approximately normal. In Figure 5, however, the distributions of z are markedly skewed, particularly in the graph in the lower right of the figure.

When looking at the results in Table 17, the impact of the *r* to *z* transformation on the RMSE's becomes clear. First, the L-W method overestimates the mean value. However, the standard deviations of the L-W means are similar to both those in the V-H method and the S-H method, indicating that the average variance of the estimate is not much different. It is the root-mean-square errors that are so much larger. This is not surprising because the back-transformed average of the theoretical L-W mean function is higher than the true rho means that are used for the RMSE calculation.

Despite all of these potential issues, all three methods provide similar estimates of the population values on average. This gives some clue as to how each method would work if a researcher were unaware of an existing moderator and just ran a meta-analysis. Each of the methods produces a fairly reasonable estimate of the average population mean. However, those conditions with small numbers of studies (*k*) still have the highest amount of variance. As is true generally in parameter estimation, researchers should always try to use large numbers of data points, in this case, numbers of studies. This is especially crucial if any type of moderator may be present.

REVCs. Another type of error was also added into these part two results. This was the random error component as derived from a normal distribution with a mean of zero and a standard deviation of .03. The Vacha-Haase method makes no attempt to estimate or correct for random error at the population level or the level of the infinite-sample effect sizes. However, both the Schmidt and Hunter and the Lipsey and Wilson methods compute an estimate of the random effects variance component (REVC). Each method uses a different computation of the REVC. Tables 18 and 19 show the results of the S-H and the L-W estimations of the REVC for each method under each condition. In Table 18, the theoretical estimate of the REVC is zero, because no random error, other than sampling error, was introduced. As expected, the estimate of the REVC for both methods is very close to zero on average. The L-W REVC is slightly higher because it is computed in *z* rather than *r* and *z* values are disproportionately higher than their corresponding *r*-values. The higher the number of studies the more closely the REVC's are to the expected zero value because sampling error is always more reduced with larger sample sizes.

When the moderator equation is added, the REVC estimates for the S-H method should approximate $.03^2 + .03^2 = .0018$. The first .03 is due to the standard deviation of the moderator and the second .03 due to the standard deviation of the random error component. Schmidt and Hunter refer to this as the total variance minus the sampling error. The moderator equation was run without incorporating sampling error to find the observed reliabilities over 10,000 times. The resulting distribution of reliabilities had a variance equal to .0018, as expected. Unfortunately, the REVC for the Lipsey and Wilson method is not directly comparable to .0018 because it is computed in *z*. Thus, there is no simple transformation of the variance in *z* that will make it directly comparable to the estimate in *r*. The L-W method uses the REVC result to recalculate the inverse variance weights.

In Table 18 the S-H and the L-W REVC's are presented for the random-effects (with moderator) condition. The REVC estimates for the S-H method are all approximately .0018, as the number of studies (*k*) grows, the REVC's also become slightly larger but the standard deviations around the estimates become smaller. The REVC for the S-H method can be expected to become larger as *k* increases because of the way in which the weighted variance of study effect sizes is computed. The method results in a biased estimate of the observed variance such that the variance estimate is too small with a small number of studies. As *k* increases, the variance estimate becomes unbiased (see Hall and Brannick, 2002).

In general, the REVCs for the L-W method are expected to increase with larger rhos but not with larger *k*s. The increase in REVC with larger rhos is demonstrated in Table 19. As expected in the .74 data the increase in *k* does not appear to have an effect on the REVC. However, there is a noticeable increase in the value of the REVC between a *k*=10 and *k*=50 in the .84 condition.

The REVC estimates in the L-W method become noticeably larger in the .84 conditions in this study. This is because when *k*=50 or higher there is a significant probability (due to the underlying distribution of *z*'s) that a very large value of *z* will be included in the analysis. For example, if one sample correlation is equal to .9999 (z = 6.10), this one *z* value will increase the estimated REVC substantially. In Figure 7, a graphic representation of this is presented. Figure 7 was constructed by choosing 1,000 randomly generated values of rho transformed to *z* versus the ln(t), where there were

1,000 randomly selected number of days between 1-35. The higher values of rs are clearly spread out from the lower values of r. If a high *z* value also has a large N (sample size) associated with it than the impact of the transformation is potentially greater, because the high *z* value is then weighted more heavily. The likelihood of that happening increases as the number of studies increases.

The initial REVC estimation in the S-H and L-W methods assumes that all of the variance beyond sampling error is random. However, if a moderator is influencing the variance, then part of the variance is not truly random. Testing for the presence of moderators thus becomes crucial in differentiating indefinable random variance from moderator variance.

It is important to remember however, that random error at the infinite effect size level is error that we cannot yet explain but that is important nonetheless. It is the quest of the researcher to try to account for and explain all variance in a true score. In the random effects model, however, there is no effort to explain part of the variability. Moderators are used to explain part of the variance; what is left over is said to be random. Thus, a mixed-effects approach, like the L-W method, is often favored.

Regression models. The Vacha-Haase and Lipsey and Wilson methods use regression models to test for the presence of a moderator. The Vacha-Haase method computes ordinary least squares regression with unit weighting. However, it also incorporates *N* (sample size) into the regression equation as a potential moderator.

The Lipsey and Wilson method uses a weighted least squares regression model. The L-W method incorporates the recalculated inverse variance weights as the weights in the procedure when no moderator is expected. However, when a moderator is suspected, L-W first runs a weighted regression that computes a revised REVC based on the residuals. Then the inverse variance weights are recalculated using this better estimate of the REVC and a second weighted regression is run. The results of this second regression are the reported results.

In order to more directly compare the V-H and L-W results, two additional regression models were computed. The straight OLS regression was run exactly as the V-H method, but without incorporating N into the moderator estimation. By removing N as a factor, the results are more similar to the L-W method. The L-W method however, is computed in *z* and then backtransformed into *r* and is indicated as WLS(z). This transformation makes the results of the L-W method incomparable to the OLS model. So the *r* to *z* transformation was also removed in one of the weighted least squares regression models indicated in Table 20 as WLS(r).

All of the regressions were computed using the natural log of time in days (1-35) rather than raw time in days, in order to satisfy the assumptions of linear regression.

Table 20 shows the results for the OLS, the V-H, WLS(z) and WLS(r) in terms of the slope estimates. In Equation 2, the slope is -.04. Therefore, the slope estimates should approximately -.04, with the exception of WLS(z) where the slope estimates should be larger due to the *r* to *z* transformation.

Table 20 shows that the slope estimates are equivalent in the methods using *r*-values; they all result, after rounding, in a slope of -.04. All three methods provide reasonable estimates of the relationship between rho and the log of time on average.

The OLS, V-H and WLS(r) results do have some slight differences in the standard deviations of the slope and the RMSE's of the slope estimates. In the *k*=10 conditions, the standard deviations and the RMSE's are different between the methods by a factor of .001. The OLS model consistently has the lowest SD's and RMSE's in the *k*=10 rows, although in the .74 condition the WLS(r) is equivalent. These results are somewhat puzzling because typically weighted least square regression would have superior results over a unit weighted procedure like OLS. It is possible that with reliabilities, the sampling error is just too small to create these types of differences. In this model in particular, having a mean N of 125 with an SD of 25 (and a minimum value of 50) may have been too high to bring out significant sampling error differences. If the average N had been lower or had N been more variable, there might have been more impact when weighting by N.

For the WLS(z) method, the slope estimates are computed in *z*. This means that in *z*'s the slope of -.04 no longer applies. Because the *z*s have a curvilinear relationship with ln(t), the slope is dependent on the number of points used in the regression. Thus, to arrive at an estimate of the slopes for each population, a regression was done on the transformed z values corresponding to the 35 time intervals of test-retest with no random error or sampling error added. These estimates of the slopes for means of .74 and .84 were -.10 and -.17. Because these are just estimates, they are not directly comparable to the results in the rest of the table, but they give an idea of how well WLS(z) estimated the slope in *z*. The RMSE's for the WLS (z) in Table 20 are computed using those numbers.

The SDs and the RMSEs in the .84 conditions follow a pattern similar to the in the results for the mean *r* =. 74 conditions. The *k*=10 condition again provides the largest values of of SDs and RMSEs. However something very unique happens in the mean=.84 conditions. The *k*=10 slope estimate is the one that matches the estimated -.17

slope most closely, however the SD and RMSE are very large. In the *k*=50 and *k*=100 rows, the slope estimates are further away from the -.17, but the standard deviations and RMSE's get smaller. The slope estimate of -.17 may not be exact because of the curvilinear shape of the *z* vs. ln(t) plot. In fact, -.19 may more accurately estimate the slope as rho becomes larger and more data points are incorporated. It indicates that as rho approaches 1.0, WLS using *z* estimates becomes less accurate. This is because as more data points are incorporated there will be a higher chance that larger values of *z* will be incorporated and the slope will get steeper.

Type I and Type II error rates. Type I and Type II error estimates are provided for all of the regression models. This is a way to directly compare all of the regression models using the same parameters. Table 21 shows the estimated Type I errors for each of the regression methods. In this study, Type I error represents the number of times that a relationship between rho and the moderator is found by chance, when the relationship does not exist. The Type I error estimates were derived as follows; for each mean rho (.74, .84) and number of studies (k), *k*-studies were generated with sampling error and matched with a random test-retest interval. A regression was done with the *k* studies in which the estimate of *r* was the dependent variable and the test-retest interval (1-35 days) was the independent variable. The regression slope was estimated and tested for significance. This process was repeated ten thousand times and the number of times the probability of the slope was less than .05 was counted. This count divided by ten thousand was the reported Type I error rate as a percentage value. This was done for each of the regression methods. In the case of the V-H regression, a random N was also matched to each of the *k*-studies, because V-H uses N in the regression model as a variable.

In OLS regression, Type I error is known to have an exact value of .05 at alpha=. 05. Using a similar line of reasoning, the other Type I errors should also be around .05.

Peculiar to this study, in the condition in which *k*=10 some of the rho's are very large values when converted to *z*. If you have a few high values in *z*s, by chance, it will look like a significant relationship is present based on the limited number of data points. This may explain why the *k*=10 conditions in the weighted least squares regression in *z* has a high Type I error rate that becomes reduced with larger numbers of studies.

The Type I error rates were in the range of the expected .05 value, although in the *k*=10 conditions, the WLS (z) method produced an excessive number of Type I errors. The overall conclusion is that all of the methods have the expected Type I error rate of about .05 in when *k* is equal to 50 or more studies.

Type II errors represent the number of times the regression fails to find the moderator. Type II errors have an inverse relationship to the Type I errors. The random error component that was added to the moderator equation (with a mean of zero and a standard deviation of .03) was intended to create some 'noise' in the moderator function. Table 22 presents the Type II error rates for each method.

All of the methods have much higher Type II error rates in the lower *k* conditions. As predicted, the added random error component 'hides' the moderator almost 50% of the time when the number of studies is small. The WLS(z) method actually proved to have the lowest Type II errors even in the low *k* conditions. The OLS and WLS(r) results are almost directly equivalent. The Vacha-Haase Type II error rates are consistently higher than all the others. However, once the number of studies is larger than fifty, all of the methods found the moderator relationship 100% of the time.

Discussion

This study set out to test methods of meta-analysis commonly used in the literature today. These methods have historically been used to analyze validity data. However, in 1998, Vacha-Haase published a groundbreaking study that used these methods to analyze reliability data. Vacha-Haase recommended the use of meta-analytic techniques to address a common reliability reporting error in the literature. Research on the misreporting of reliability coefficients has shown that as many as one-half of researchers do not report the appropriate reliability coefficient for their study (Vacha-Haase et. al., 1999; Whittington, 1998). Meta-analysis can be used to evaluate how reliability will function across conditions, thereby allowing researchers to predict how reliability will behave in their local populations. Thus, for studies in which reliability is not reported or is misreported, meta-analysis of reliability might be used as a suitable alternative.

In addition, very little research has been done to discover the impact of moderators on reliability coefficients. Meta-analysis in combination with a regression technique is a solid methodological approach to deciding whether moderators explain variance in effect sizes. However, the application of both meta-analytic and regression techniques in reference to reliability coefficients has not been well studied.

This study sought to address the question of which meta-analytic approach is the best one to use for reliability coefficients. The question was investigated in two conditions, one in which moderators were absent (the fixed-effects case), and one in which a moderator was present (the mixed- or random-effects case). The three methods of meta-analysis selected for study included the methods outlined by Vacha-Haase (1998), the Hunter and Schmidt (1990) "bare-bones" meta-analytic technique and the Lipsey and Wilson (2001) version of the 'random-effects' meta-analytic model developed by Hedges and colleagues. These methods were selected because they either were designed for the analysis of reliability data (Vacha-Haase, 1998) or because they are methods that are commonly used and believed to be widely applicable (Hedges & Vevea, 1998; Hunter & Schmidt, 1990) and therefore likely to be applied to the meta-analysis of reliability data.

A Monte-Carlo technique allowed for the setting of known population parameters against which the performance of each of the three models could be judged. Each of the models was used to estimate the mean and (except for Vacha-Haase) the random-effects variance component in both fixed- and random-effects conditions.

Time between test and retest was simulated as a moderator of the underlying reliability. Two regression models (V-H and L-W) were fit to meta-analytic data to see how they compared in recovering a known parameter. In addition, new methods of data analysis were studied (unit and sample size weighted regression in *r*) in order to better understand the reasons for the differences between the H-V and L-W models. The new methods helped to disentangle the effects of the meta-analytic weights and the effects of the *r* to *z* transformation.

Part One

Estimates of mean and variance in a no-moderator, fixed effects condition. In Part One of this study, each of the methods was computed for a no-moderator situation in which the only source of variance in observed reliability estimates is sampling error. The true population reliability coefficients were set to .74 and .84. The results of this analysis showed that the Lipsey and Wilson method consistently overestimated the true reliability. On the other hand, compared to the other two methods, the L-W method had a somewhat smaller standard deviation and root-mean square error (RMSE), especially when the number of studies used in the meta-analysis was small. The Vacha-Haase and Hunter and Schmidt methods tended to underestimate the true reliability values, and the standard deviation estimates were about .001 larger in magnitude than the L-W results.

Overall, the results suggest that the L-W method was somewhat better at estimating the population reliability when no moderator was present. The advantage for the L-W method was most evident when the number of studies used in the meta-analysis was small (ten). Once the number of studies used was fifty or more, the differences among the methods were negligible.

The L-W method had the best performance under the fixed-effects condition. The V-H and the S-H methods sometimes estimated the mean equally as well at the L-W, but the L-W method never did worse and most of the time did better at correctly estimating the mean effect size. However, as Hunter and Schmidt (1990) have argued, fixed effects scenarios are rarely plausible in actual data because of measurement error and other artifacts that produce variance in addition to that produced by sampling error.

The underestimation of the true rho values by V-H and S-H methods is explained by the skewness of the sampling distribution of the reliability coefficient. The negative skewness of the distribution causes the arithmetic mean to underestimate the true mean. This is because random individual study values lower than the true population mean are likely to be farther away from that mean than those study values that are higher than the true population mean due to the negative skew. This explains why the estimates of the $\overline{\rho}$ in V-H and S-H results are underestimates of the true mean. As for the L-W results, the overestimation of $\overline{\rho}$ is primarily due to the *r* to *z* transformation that normalizes the distribution but creates larger values of rho when backtransformed.

Random Effects Variance Components (REVCs). The random-effects variance components were calculated for both the S-H and L-W models, although the computations are different. The S-H REVC is based on the total variance minus the estimated sampling error variance. The L-W variance is based on the chi-square distribution, and compares the observed sum of squared deviations to the expected sum of squares. In part one, the S-H REVC is close to zero because only sampling error is included in the estimates of rho. The L-W REVCs are higher for part one, but this is mostly due to fact that the REVC is calculated using *z* in the L-W method. In both methods, estimates of the REVC that are less than zero are set to zero. This results in the positive bias of the estimated REVC shown in Table 4.

Part Two: Analysis with the Introduction of the Moderator

Mean and variance. In part two, a moderator function was used to simulate effect sizes that vary across conditions. The moderator used was a ‘real-world’ function modeling time decay in test-retest reliabilities. Two means were used with the same function: the higher mean ($\overline{\rho} = .84$) simulating the cognitive ability tests and the smaller mean ($\overline{\rho} = .74$) simulating job satisfaction measures. When the moderator was added, the previously negatively skewed sampling distribution of reliability (in the no moderator situation) now became positively skewed (see Figures 5 and 6). A random error component was also added at the population level so that even after the moderator was accounted for, there was still a positive REVC. Samples were drawn from the populations, so the observed distributions of reliability coefficient showed variability due to the combined effects of the moderator, the sampling error and the random error term.

Because the moderator introduced another type of variance, standard deviations and RMSEs were larger than in part one. This was expected. However, the pattern of results in Part Two is very different from that in part one.

For the V-H and S-H methods, the estimated means, SDs and RMSEs were very similar. The two methods estimated the grand mean reliability ($\overline{\rho}$) within .005 in every condition. As was expected, due to sampling error and random error, the methods had much higher SDs and RMSEs when *k* (number of studies) was equal to 10.

The Lipsey and Wilson method lost its advantage in estimating the reliability coefficients once the moderator was added. The L-W method continued to overestimate the population mean; however in this condition it had higher SDs and RMSEs than either of the other two methods. This pattern was especially apparent in the *k*=10 conditions and more so in the $\overline{\rho}$ =.84 condition. This is due primarily to the inclusion of the *r* to *z* transformation. Many researchers have argued for the inclusion of the *r* to *z* transformation (James et. al, 1986), and it seems that in the no-moderator condition, the transformation enhances the outcome. However, once the moderator was added, the underlying distribution of rhos was *positively* skewed by the transformation.

The impact of the r ***to*** z ***transformation in a moderator condition.*** The positive skew in the with-moderator distribution is magnified when the *r* to *z* transformation is applied in the L-W method. This is due to the fact that as *r*s get larger, the corresponding *z*s are disproportionately larger (that is, *r* to *z* is a nonlinear transformation). As an example, when rho is. 99, the corresponding *z* is 2.65, however when rho is .9999, the corresponding *z* is 6.10. This shows that when *r* is large, large changes in *z* occur in response to very small changes in *r*. The net effect in the rho-moderator relationship is evident in Figure 7. Figure 7 is a graphic representation of 1,000 randomly generated values using the modeled moderator function, transformed to *z* with the Fisher *r* to *z*, then plotted against the corresponding ln(t). The rapidly increasing *z* values transform a linear relationship into a nonlinear one. As mentioned previously, most meta-analytic techniques have been developed and used for the study of validity, where effect sizes tend to be small. Reliability estimates, however, tend to represent rather large effect sizes (many are greater than .90). Thus, the *r* to *z* transformation can be expected to introduce more variance to the distribution of reliability estimates than to a distribution of validity estimates. This may serve as a cautionary flag for researchers. When estimating reliability coefficients, particularly when expected reliability values are in the upper range, researchers should be aware of those conditions where a moderator might be present. If confronted with such a situation, use of the *r* to *z* transformation should be weighed against the changes that may occur both in the distribution and in the underlying moderator relationships.

REVC. In part two, the REVCs for the S-H and L-W methods were calculated. The REVCs for both methods increased in value as expected in the presence of a moderator and random error. The S-H method slightly overestimated the REVC on average.

In general, the REVC for the L-W method was expected to increase with larger $\overline{\rho}$, but not with larger *k*. Results consistent with this expectation can be seen in Table 5. Such a result can be explained by the *r* to *z* transformation. As mean *z* becomes larger, the distribution also becomes more variable. Note, however, that whereas in the .74 condition as *k* increases the REVC remains essentially unaffected, in the .84 condition an increase in the value of the REVC is observed between the *k* of 10 and the *k* of 50. This result appears due to the probability that a very large z-value will be included in the analysis. Recall that there were only 35 accepted population time values (t= 1 to 35), and they were *uniformly distributed*. Thus the likelihood of a value =.95 (maximum) is equal to that of any other value and will have a 1 in 35 chance of occurring in the sample of studies. When random error and sampling error are added, this value could approach .9999, which was the cutoff for this study. This corresponds to a z value of 6.10. Hence, though not specifically tested in this study, one can predict that at a *k* of 35 or greater, on average at least one large z-value is being used in the analysis. This outcome also indicates that if a researcher is using the *r* to *z* transformation with a moderator present, then REVC estimates may be adversely impacted by the transformation, especially as *k* (number of studies) increases.

Moderator analysis. A second purpose for meta-analyzing reliability coefficients, according to Vacha-Haase (1998), is to identify moderators of reliability. Reliability is defined as "the consistency with which individuals are rank ordered by measurement across parallel forms, repeated measures or other estimates of consistency in measurement" (Spearman, 1910, p. 272). Thus, a moderator can be any factor that would impact the consistency of measurement. In the case of test-retest reliability, the amount of time delay between the first test and the second can create significant changes in the scores. This is a fairly obvious moderator, but other factors such as gender, race, education level, amount of sleep the night before the test, personality, and many others can influence the consistency of scores.

Regression is commonly used to seek out the presence of moderators. The Vacha-Haase and the Lipsey and Wilson methods both outline regression methods for the detection of moderators. The V-H regression is based on the ordinary least squares method, and can include multiple moderators. In this study, the method was used to estimate the impact of the logarithm of time between test-retest and the impact of sample size (N). V-H does not use any study or effect-size weights in the regression analysis because N is included as a potential moderator. Lipsey and Wilson on the other hand used inverse variance weights in a weighted least squares regression model. In the L-W method, when the effect size estimates are correlations, the *r* to *z* transformation is applied, then the effect-size weights become N_i-3 (three less than the sample size). This is because the expected sampling variance of a *z*-transformed correlation is (1/(N-3)). To better understand any differences in results for the V-H and L-W methods, two other regressions were computed. The OLS (unit weights) showed the effect of computing a regression without sample size as an independent variable. The difference between OLS and V-H is solely that V-H includes N as an independent variable. The WLS regressions in *r* used N_i-3 as the study weight. The difference between this model and the L-W model is solely the *r* to *z* transformation.

Slope estimates. The first set of results from the regressions was the slope estimates. The parameter was $\beta =$ -.04 and the slope estimates from each of the regression models computed on *r* (OLS, V-H and WLS(r)) should have accurately estimated this slope. SDs and RMSEs were computed for each slope estimator as well. All three of the methods computed in *r* estimated the slope to be -.04 on average. The WLS(r) and the unit weighted OLS had SDs and RMSEs that were almost equivalent (see Table 6). This is a little puzzling because a WLS procedure should have better estimates due to the correction for sampling error. However, it appears that reliability estimates are in the range where sampling error is very small. The sampling error estimate in the Schmidt and Hunter model supports this idea. In that equation, as the effect size statistic approaches one, the sampling error variance approaches zero. Thus, weighted regression may not have much of a unique predictive value over and above a unit-weighted procedure when reliability is the effect size of interest.

The *r* to *z* transformation is presented in the WLS (z) results. The slope estimate is different because when the *r*s are converted to *z*s, the values become much higher. The slope estimates are therefore reported as they relate to the *z* values. The best linear estimates of the slope in the .74 conditions would be around -.10 and in the .84

conditions would be around -.17. However, as has been previously discussed, the *r* to *z* transformation creates a curvilinear relationship between z values and ln(time). This means that the slope estimates will change depending on the number of *z* values that are in the highest ranges. This effect appears to be the reason that the reported slope estimates in the .84 conditions change as *k* (number of studies) becomes larger.

Type I and Type II error rates. For OLS, the Type I error rate at alpha=.05 is known to be an exact value of .05. Thus, the Type I error estimates are in general expected to be approximately .05 or 5% across methods. For all of the methods, with the exception of the Lipsey and Wilson WLS(z) method, the empirical estimates of Type I errors were close to .05.

The WLS(z) method, however, produced values that are much greater than the expected 5% in the *k*=10 conditions. This is most likely due to the chance presence of very high values of *z* that will result in large slope estimates that are mistakenly judged significant. This is yet another concern for the *r* to *z* transformation that has been exposed by this study, in the case where a moderator is present.

A Type II error occurs when a moderator is present, but the regression slope is not significant and thus there is a failure to detect a real moderator. Type II error is related to the power to detect the moderator. Those methods that can identify the real or true moderator most often (lower Type II error) are said to have higher power.

In this study all of the methods have much higher Type II error rates in the *k*=10 conditions. This is not surprising because with a small number of studies, the random error and sampling error are more likely to mask the moderator variance.

Type II error rates are only directly comparable when the Type I error rates are equal. If all Type I error rates are .05, then we should prefer the method that produced the fewest Type II errors. The Lipsey and Wilson WLS(*z*) method had the lowest Type II error rates when the number of studies was low (*k*=10). In isolation, this result would be encouraging for the *r* to *z* transformation. Unfortunately, the power to detect the moderator comes at the cost of having a higher Type I error rate, and thus the comparison and choice among the methods is not a clear as one would like.

The power of the regression slope estimate appears to pass .90 somewhere between 10 and 50 studies for the simulated reliability data considered in this paper. Thus, the power for detecting moderators in reliability data may be surprisingly good.

The choice of r *versus* z. There is something of a debate in the literature regarding whether to analyze the correlation effect size in *r* or *z* (Erez, Bloom and Wells, 1996; Hunter and Schmidt, 1990; Silver and Dunlop, 1987; Hedges and Olkin, 1985). According to the current results, when the population has a single value (the fixed-effects case), the transformation appears to normalize the sampling distribution and results in better estimates of the population value than does the untransformed *r*. Therefore, *z* appears preferable to *r* for a meta-analysis in the fixed-effects case.

When the population rho is a random variable (the random-effects case), the advantage of the transformation disappears. The effect of the transformation is to skew the distribution of rho so that the estimate of the mean becomes biased. The random-effects variance component is expressed in *z*, which is a problem because it cannot be directly converted to *r*, the original unit. Rather, the REVC must be used in an equation to make a prediction of some sort, and the predicted value of *z* must be back transformed to *r* for interpretation. Therefore, *r* appears to be a better choice than *z* for a random-effects meta-analysis in which the main goal is to estimate the mean and REVC for a set of studies.

The choice of *r* or *z* becomes more complicated when moderators are considered. Unlike ordinary regression, in meta-analysis there is heteroscedasticity inherent in the data because the studies have different sample sizes, and thus different amounts of sampling error associated with them. If the studies can be considered a random sample (if sample size is not correlated with effect size) then heteroscedasticity may not be a large problem in interpreting the results of the moderator analysis. Weighted regression seems to be an appropriate way to incorporate the impact of sampling error into the analysis, and this can be done in either *r* or *z*.

The current study showed additional problems in using *z* for moderator analysis as well as an advantage of doing so. First, if the moderator is linearly related to the size of *r*, then it will be nonlinearly related to the size of *z*, and vice versa. A potential solution to this problem might be polynomial regression. Second, if there is an additional error term beyond the moderator at the infinite-sample effect size level, and this term is homogeneous in *r*, it will be heterogeneous in *z*. Figure 7 shows both problems. The implication is that it would be difficult to position confidence intervals around the regression line computed in *z*. A third difficulty is that the slope in *z* changes as the mean *z* changes because of the nonlinear transformation. Thus it will be difficult to interpret the slope of a moderator computed in *z*. Finally, we have the inflated Type I error rate when the number of studies is small. All these problems argue for the analysis in *r* and against the analysis in *z*.

The advantage to using *z* according to the current study is the greater power of the test for the presence of the moderator. When the number of studies is small, the advantage is somewhat mitigated by the inflated Type I error rate.

Study limits. The purpose of this study was to look at only three different meta-analytic techniques and their application to reliability coefficients in a very controlled context. Thus, the study shares some of the limitations inherent in the use of the Vacha-Haase, Schmidt and Hunter and Lipsey and Wilson methods. There are many other types of meta-analysis that could be evaluated, however the current three methods were chosen based on their popularity of usage and because they had some interesting differences from one another.

Two types of regression techniques, OLS and WLS were evaluated. However, the regressions were run in such a fashion as to disentangle the effects due to both weighting and the *r* to *z* transformation. In an effort to focus on those factors and provide for a direct comparison of results, the weighted least squares regression in *r* was done using the same weights as the WLS in *z*. It is a limitation of this study that the inverse variance weights normally applied to WLS when using *r*s were not calculated. This may be part of the reason (in addition to small sampling variance of reliability coefficients) that the WLS results did not outperform the OLS results, as they would normally be expected to (Steel and Kammeyer-Muller, 2002).

In this study, the impact of the number of days between test and retest was used as a moderator. Although this moderator was taken directly from a real world test in the .84 conditions, it only served as an estimate of what might happen in a job satisfaction or similar measure in the .74 conditions. Furthermore, this (log transformed) moderator had a linear relationship with the reliability estimates. In reality, the moderator may not have a perfectly linear relationship with the effect size statistic.

Three levels of *k* (number of studies), 10, 50 and 100, were used in this study. This provided only a limited view of how the meta-analytic methods were functioning when there were smaller numbers of studies. Based on the current results, gathering additional data between 10 studies and 50 studies is warranted to better understand the Type I and Type II error rates of the regression techniques.

As expected based on previous research (James et. al, 1986, p. 446), the distribution of *r* was negatively skewed in the fixed effects condition. However, the distribution became positively skewed with the addition of the moderator and random error in this study. This may be a unique feature of moderator used. The degree and direction of the skewness in other *r* distributions may be very different with other moderator variables.

Future research. This study brought to light some interesting ramifications of using the *r* to *z* transformation when moderators are present. Research should be conducted to determine whether polynomial regression or some other analysis might prove to be a better estimator when using *z* for moderator analysis. This research could help clarify why the analysis in *z* had better power than the analyses in *r* when using the WLS method of regression.

In this study there were no additional levels between *k*=10 and *k*=50 studies. Overall, the larger standard deviations for all three methods in the *k*=10 conditions highlight the need for caution when there are smaller numbers of studies being studied. In these conditions the mean effect sizes were off by as much as .002 from the true mean.

Additional research is suggested to determine what happens to the SD's when the number of studies is increased to some number between 10 and 50.

It appears that because reliabilities are generally fairly large (> .70), more attention should be paid to the size of sampling error estimates as reliability estimates become larger. This is especially valid information when using regression methods to search out moderators. If the sampling error is very small at larger values of reliability, the differences between methods that weight for sampling error and those that don't are reduced. The sampling variance of the correlation is approximately:

$$\sigma_e^2 = \frac{(1-\rho^2)^2}{N}$$

Using this formula, when reliability is .64, sampling variance is estimated to be .003 with an N of 125. When reliability is .74, that figure is reduced to .002, at .84 it becomes .001, and at .94 it becomes .0001. Further research is necessary to determine exactly how small the sampling error typically is within the range of common reliabilities.

Conclusions

This study aimed to find the best meta-analysis method for reliability coefficients. The results have provided several conclusions and contributions to the literature.

First, when no moderator is present (fixed condition), the three meta-analytic methods were almost equally good at estimating the true population ρ_i. However, the Lipsey and Wilson method had a consistent advantage over the other methods, which was more pronounced when the number of studies was small. Thus the L-W method is recommended for use when the required meta-analysis is for fixed effects.

Second, once a moderator produces variance in reliability coefficients, the Lipsey and Wilson method becomes significantly less accurate due to the *r* to *z* transformation and the method begins to consistently overestimate the true population mean effect size value. In the presence of a moderator like the one in this study, the Vacha-Haase and the Schmidt and Hunter methods appear equally good at estimating the population effect size and are better estimators of the mean than is the L-W method. The Schmidt and Hunter method is more highly recommended because it estimates the random effects variance component in addition to the mean and thus provides more information to the researcher.

Third, when using regression to evaluate a moderator, weighted least squares regression is usually more powerful than using a unit weighted ordinary least squares method (Steel and Kammeyer-Muller, 2002). This is because the weighted least squares methods use an estimate of sample size to weight the regression and to reduce the impact of sampling error in the prediction. Even though the sampling error associated with reliability may be small, correcting for it within the regression still produces a better estimate of the slope. Thus, based on current information, computing WLS regression in *r* appears the best method to test for moderators in reliability studies.

In conclusion, a new and somewhat unique combination of methods is recommended. Because most real world situations do include moderators, researchers should apply the Schmidt and Hunter technique for meta-analysisto obtain the best estimates of the overall mean and random-effects variance component. Researchers who are also interested in evaluating continuous moderators of reliability should compute a weighted least squares regression in *r*, to obtain the best estimate of the slope.

References

Algera, J.A., Jansen, P.G., Roe, R.A., & Vijn, P. (1984). Validity generalization: Some critical remarks on the Schmidt-Hunter procedure. *Journal of Occupational Psychology, 57,* 197-210.

American Psychological Association. (1994). *Publication manual of the American*

Psychological Association (4th ed.). Washingtion, DC: Author.

Bem, S. L. (1981). Bem Sex-Role Inventory: Professional manual. Palo Alto, CA: Consulting Psychologist Press.

Caruso, J.C., Witkiewitz, K., Belcourt-Dittloff, A. & Gottlieb, J.D. (2001). Reliability of scores from the eysenck personality questionnaire: A reliability generalization study. *Educational and Psychological Measurement, 61*, 675-689.

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

Erez, A., Bloom, M.C. & Wells, M.T. (1996). Using random rather than fixed effects models in meta-analysis: Implications for situation specificity and validity generalization. *Personnel Psychology, 49*, 275-306.

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

Hall, S.M. (2000). A comparison of the Schmidt and Hunter method of meta-analysis and the random-effects method of meta-analysis under "Real World" conditions. Unpublished doctoral dissertation, University of South Florida, Florida.

Hall, S.M. & Brannick, M.T. (2002). Comparisons of two random-effects methods of meta-analysis. *Journal of Applied Psychology, 87,* 377-389.

Hedges, L.V., & Olkin, L. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

Hedges, L.V. & Vevea, J.L. (1998). Fixed and random-effects models in meta-analysis. *Psychological Methods, 3,* 486-504.

Hogan, T.P., Benjamin, A. & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement, 60,* 523-531.

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

Lawrence, R.J., Demaree, R.G & Mulaik, S.A. (1986). A note on validity generalization procedures. *Journal of Applied Psychology, 71*(3), 440-450.

Lipsey, M.W. & Wilson, D.B (2001). Applied Social Research Methods Series, 49. *Practical meta-analysis.* Newbury Park, CA: Sage.

Nunnally, J.C. (1978). *Psychometric theory (2nd ed.).* New York: McGraw-Hill.

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

Raju, N.S and Drasgow, F. (in press). Maximum likelihood estimation in validity generalization. In K. Murphy (ed.) *Validity generalization: A critical review.*

Rosenthal, R. (1987). Meta-analytic procedures for social research (Rev. ed.). Newbury Park, CA: Sage.

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

Schmidt, F.L. & Hunter, J.E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62,* 529-549.

Schmidt, F.L., Hunter, J.E., & Raju, N.S. (1988). Validity generalization and situational specificity: A second look at the 75% rule and Fisher's r to z transformation. *Journal of Applied Psychology, 73*, 665-672.

Schmitt, N. & Noe, R.A. (1986). On shifting standards for conclusions regarding validity generalization. *Personnel Psychology, 39,* 849-851.

Silver, N. & Dunlop, W. (1987). Averaging coefficients: Should Fisher's *z*-transformation be used? *Journal of Applied Psychology, 72,* 3-9.

Spearman, C.E. (1910). Correlation calculated from faulty data. *British Journal of Pscyhology, 3,* 271-295.

Spector, P.E. & Levine, E.L. (1987). Meta-analysis for integrating study outcomes: A Monte Carlo study of its susceptibility to Type I and Type II errors. *Journal of Applied Psychology, 72,* 3-9.

Steele, P. D. & Kammeyer-Muller, J.D.(2002). Comparing meta-analytic moderator estimation techniques under realistic conditions. *Journal of Applied Psychology, 87,* 96-111.

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

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.R. & Thompson, B. (2000). Sample compositions and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement, 60,* 509-522.

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

Viswesvaran, C. & Ones, D. S. (2000). Measurement error in “Big Five Factors” personality assessment: Reliability generalization across studies and measures. *Educational and Psychological Measurement, 60*, 224-235.

Viswesvaran, C., Ones, D.S. & Schmidt, F.L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81,* 557-574.

Whitener, E. M. (1990). Confusion of confidence intervals and credibility intervals in meta-analysis. *Journal of Applied Psychology, 75*, 315-321.

Whittington, D. (1998). How well do researchers report their measures? An evaluation of measurement in published educational research. *Educational and Psychological Measurement, 58,* 21-37.

Yin, P., & Fan, X. (2000). Assessing the reliability of the Beck Depression Inventory scores: Reliability generalization across studies. *Educational and Psychological Measurement, 60,* 201-233.

Table 1

Overview of the three meta-analytic methods

Table 1: Comparison of Methods	**Vacha-Haase**	**Hunter and Schmidt**	**Lipsey and Wilson**
1.Weight effect size statistics to find the average effect size across studies.	Vacha-Haase uses a unit weighted average of the reliability.	Hunter and Schmidt weight each reliability statistic by it's sample size (N). Then they find the average weighted reliability.	Lipsey and Wilson suggest using an inverse variance weight. Because they use the Fisher's r to z transformation, they calculate the inverse variance weight to be (N-3) for each reliability. Next they average the inverse variance weighted statistics.
2. Compute the Variance of the observed effect sizes.	Vacha-Haase computes a unit-weighted variance. She describes the distribution using Box and Whisker plots.	Calculate the weighted (N) variance of the statistic across studies.	Calculate the weighted (N-3) variance.
3. Correct for sampling error.	Vacha-Haase includes sample size in the moderator analysis, but does not suggest any corrections when the sample size does account for significant variance. Vacha-Haase proposes a fixed-effects model.	Correct the variance by subtracting the amount attributed to sampling error. Using $\sigma_{\mathrm{e}}^2 = \left(1-\bar{r}^2\right)^2 / \left(\bar{N}-1\right)$ to estimate variance due to sampling error and subtract from amount of variance observed across all studies.	Estimate the random-effects variance component through a procedure analogous (but not identical) to the Hunter and Schmidt method. If the random effects variance component is greater than zero, re-estimate the value of the mean with new weights.
4. Corrections for other artifacts [take out this row. No other corrections in this study.]	Vacha-Haase does not address artifact corrections.	Hunter and Schmidt have a long list of artifacts for meta-analysis of test validation studies. There are no specific descriptions of how these corrections would apply to a meta-analysis of reliability.	Lipsey and Wilson describe corrections for single artifacts, but do not describe how such corrections would apply to the meta-analysis of reliability.
4. Decide whether moderators are present.	Vacha-Haase suggests thinking of all conceivable moderators, then developing a coding system to code each moderator into a variable. Assume moderators are present.	Hunter and Schmidt suspect moderators only when V_{θ} is large.	Test for the homogeneity of effect sizes.
5. Estimate moderator effects.	Perform unweighted least squares regression analyses to explore how well the coded study features predict variations in the reliability coefficients.	The moderator analysis proposed by Hunter and Schmidt (1990) suggested a series of meta-analyses, where effect sizes were divided into groups based on moderators and then each group was meta-analyzed independently.	If homogeneity is rejected, then a test for moderators is performed. Lipsey and Wilson suggest a weighted regression analysis.

Table 2

Sample data used for the examples of how each method works

Study	r_i	N	Test-Retest Interval in Days
1	0.88	85	14
2	0.95	84	3
3	0.85	56	21
4	0.9	70	14
5	0.6	45	90
6	0.4	32	180

Table 3

Confidence intervals for Vacha-Haase sample data

LOWER	MEAN	UPPPER LIMIT
0.59	0.76	0.94

Table 4

Regression output of the V-H Ordinary Least Squares regression analysis of sample data

SUMMARY OUTPUT

Regression Statistics					
Multiple R	0.99				
R Square	0.98				
Adjusted R Square	0.97				
Standard Error	0.03				
Observations	6				
ANOVA					
	df	SS	MS	F	Significance F
Regression	2	0.23	0.11	103.13	0.0017
Residual	3	0.00	0.00		
Total	5	0.23			
	Standardized Coefficients	Standard Error	t Stat	P-value	
Intercept	0	0.11	7.06	0.01	
Interval days	-.849	0.00	-5.81	0.01	
N	.16	0.00	1.09	0.35	

Table 5

Data calculations for the Schmidt and Hunter method using sample data

Study	r	N	Time Interval	N*r	r- $\bar{r}$	(r- $\bar{r}$)^2	N*(r- $\bar{r}$)^2
1	0.88	85	14	12.32	0.06011	0.003613	0.307098
2	0.95	84	3	2.85	0.13011	0.016928	1.421949
3	0.85	56	21	17.85	0.03011	0.000906	0.050762
4	0.9	70	14	12.6	0.08011	0.006417	0.449205
5	0.6	45	90	54	-0.21989	0.048353	2.175871
6	0.4	32	180	72	-0.41989	0.17631	5.64191
$\sum$		372		305			10.0468
Weighted r	0.81989						

Table 6

Credibility interval for the S-H example data

LOWER LIMIT	UPPER LIMIT
0.50846	1.13132

Table 7

Approximate confidence intervals for the S-H example data

LOWER LIMIT	MEAN	UPPER LIMIT
0.69	0.82	0.95

Table 8

Lipsey and Wilson sample data and calculations

Lipsey Wilson Data								
Study	r	N	Time Interval	Fisher z	*w*	*w*z*	z^2	*w**z^2
1	0.88	85	14	1.38	82	112.8129	1.892737	155.2044
2	0.95	84	3	1.83	81	148.3742	3.355421	271.7891
3	0.85	56	21	1.26	53	66.5761	1.57792	83.62975
4	0.9	70	14	1.47	67	98.63871	2.16743	145.2178
5	0.6	45	90	0.69	42	29.11218	0.480453	20.17903
6	0.4	32	180	0.42	29	12.28582	0.179478	5.204874
Σ		372			354	467.8	9.653439	681.225

Table 9

Confidence intervals for L-W in zs using the example data

LOWER LIMIT	MEAN	UPPER LIMIT
1.07	1.24	1.41

Table 10

Confidence intervals for the L-W method backtransformed into rs

LOWER LIMIT	**MEAN**	**UPPER LIMIT**
0.79	0.85	0.89

Table 11

WLS regression results using the example data and L-W method

```
                         SUMMARY OUTPUT

 Mean ES    R-Square      N

  1.2397       .8463      6.0000

                              ANOVA

              Q              df              p

Model        22.0943       1.0000        .0000

Residual      4.0121       4.0000        .4044

Total        26.1064       5.0000        .0001

                       REGRESSION RESULTS

                  B        SE   -95% CI     +95% CI       Z           P        Beta
CONSTANT   1.5658    .1130   1.3443   1.7872   13.8571   .0000    .0000
INTERVAL    -.0072    .0015    -.0102    -.0042   -4.7005    .0000     -.9200
```

Table 12

Approximate credibility intervals for the L-W estimates in the example data

LOWER LIMIT **CREDIBILITY**	**UPPER LIMIT** **CREDIBILITY**
0.67	0.93

Table 13

Comparison of confidence interval results across methods for the example data

VACHA-HAASE METHOD		
Lower Limit	Mean	Upper Limit
0.59	0.76	0.94
HUNTER AND SCHMIDT METHOD		
Lower Limit	Mean	Upper Limit
0.69	0.82	0.95
LIPSEY AND WILSON MIXED EFFECTS METHOD		
Lower Limit	Mean	Upper Limit
0.79	0.85	0.89

Table 14

Approximate credibility intervals between the S-H and L-W methods using the example data

LOWER CREDIBILITY LIMIT	**UPPER CREDIBILITY LIMIT**
HUNTER AND SCHMIDT	
0.51	1.13
LIPSEY AND WILSON	
0.67	0.93

Table 15

Data Summary

Population Parameters		
Part 1:	Means:	.84, .74
	Standard Deviations:	.00, .00
Part 2:		
	Average of the Means:	.84, .74
	Standard Deviations (due to presence of moderator):	.03, .03
	Random Error	Distribution with a mean of 0 and a standard deviation of .03
	Slope	-.04
	REVC (Schmidt and Hunter estimate)	.0018

Table 16

Estimates of the Mean for Fixed-Effects Conditions

		Vacha-Haase			Schmidt and Hunter			Lipsey and Wilson		
Mean ρ	Studies (*k*)	M	SD	RMSE	M	SD	RMSE	M	SD	RMSE
.74	10	.7378	.0140	.0141	.7378	.0136	.0138	.7404	.0135	.0135
.74	50	.7386	.0058	.0060	.7386	.0057	.0058	.7412	.0056	.0058
.74	100	.7384	.0041	.0044	.7385	.0040	.0042	.7412	.0039	.0041
.84	10	.8390	.0087	.0088	.8391	.0085	.0086	.8410	.0084	.0085
.84	50	.8390	.0038	.0039	.8390	.0037	.0038	.8409	.0036	.0038
.84	100	.8389	.0028	.0030	.8390	.0027	.0029	.8410	.0027	.0029

Table 17

Estimates of the Mean for Mixed (Random)-Effects Conditions

		Vacha-Haase			Schmidt and Hunter			Lipsey-Wilson		
Mean ρ	Studies (*k*)	M	SD	RMSE	M	SD	RMSE	M	SD	RMSE
.74	10	.7432	.0195	.0195	.7432	.0193	.0193	.7491	.0198	.0206
.74	50	.7435	.0086	.0086	.7435	.0087	.0087	.7500	.0088	.0109
.74	100	.7430	.0064	.0064	.7431	.0064	.0064	.7500	.0066	.0089
.84	10	.8424	.0170	.0170	.8424	.0171	.0171	.8522	.0205	.0222
.84	50	.8436	.0074	.0074	.8436	.0075	.0075	.8544	.0095	.0144
.84	100	.8438	.0054	.0054	.8438	.0054	.0054	.8546	.0068	.0131

Note. For this table, the moderator is operating to produce variance in the effect sizes, but the moderator is not analyzed in the meta-analysis. For the Lipsey-Wilson method, results were analyzed in *z*, but the reported mean, SD and RMSE values were based on *z* transformed back to *r* at the end of each of the 1,000 meta-analyses.

Table 18

Estimates of the Variance (REVC) for Part I Fixed-Effects Conditions

		Hunter-Schmidt (Total Variance -Sampling Error Estimate)		Lipsey-Wilson (V theta for Z's)	
Mean ρ	Studies (*k*)	M	SD	M	SD
.74	10	.0002	.0005	.0016	.0027
.74	50	.0001	.0002	.0007	.0011
.74	100	.0001	.0002	.0004	.0007
.84	10	.0001	.0002	.0016	.0026
.84	50	.0001	.0001	.0006	.0010
.84	100	.0000	.0001	.0005	.0007

Table 19

Estimates of the Variance (REVC) for Mixed (Random)-Effects Conditions

		Hunter-Schmidt (Total Variance -Sampling Error Estimate) ** REVC=.0018		Lipsey-Wilson (V theta for Z's)	
Mean ρ	Studies (*k*)	M	SD	M	SD
.74	10	.0018	.0016	.0054	.0060
.74	50	.0020	.0007	.0049	.0027
.74	100	.0021	.0005	.0049	.0019
.84	10	.0019	.0013	.0195	.0404
.84	50	.0021	.0006	.0252	.0367
.84	100	.0021	.0004	.0254	.0192

Table 20

Estimates of the Slope (coefficient of ln (t))

		Unit Weighted OLS			Vacha-Haase		
Mean ρ	Studies (*k*)	M	SD	RMSE	M	SD	RMSE
.74	10	-.04	.025	.025	-.04	.028	.028
.74	50	-.04	.008	.008	-.04	.008	.008
.74	100	-.04	.006	.006	-.04	.006	.006
.84	10	-.04	.020	.020	-.04	.022	.022
.84	50	-.04	.007	.007	-.04	.006	.006
.84	100	-.04	.004	.004	-.04	.004	.004
		WLS (z)			WLS (r)		
Mean ρ	Studies (*k*)	M	SD	RMsE	M	SD	RMSE
.74	10	-.10	.060	.060	-.04	.025	.025
.74	50	-.10	.022	.022	-.04	.008	.008
.74	100	-.10	.015	.015	-.04	.006	.006
.84	10	-.18	.113	.113	-.04	.021	.021
.84	50	-.19	.060	.063	-.04	.007	.007
.84	100	-.19	.041	.047	-.04	.004	.004

Table 21

Percentages of Type I Errors

Mean ρ	Studies (*k*)	% Type 1 OLS	% Type 1 Vacha-Haase	% Type 1 LW (z)	% Type 1 LW (r)
.74	10	5.06%	4.85%	7.18%	5.04%
.74	50	4.8%	4.9%	5.72%	5.11%
.74	100	4.84%	4.87%	5.07%	4.92%
.84	10	4.88%	4.78%	8.11%	4.84%
.84	50	4.99%	5.29%	5%	4.52%
.84	100	4.63%	5.21%	5.29%	4.49%

Table 22

Percentages of the Type II Errors in the Four Different Regressions

		OLS	VH	LW (z)	LW (r)
Mean ρ	Studies (*k*)	Total Percentage Type II Errors			
.74	10	57.8%	61.2%	44.5%	57.3%
.74	50	0.9%	0.8%	0.3%	0.5%
.74	100	0%	0%	0%	0%
.84	10	43.8%	49.1%	29.8%	44.1%
.84	50	0%	0%	0%	0%
.84	100	0%	0%	0%	0%

Figure 1. Observed distribution of reliability estimates based on Hogan et. al

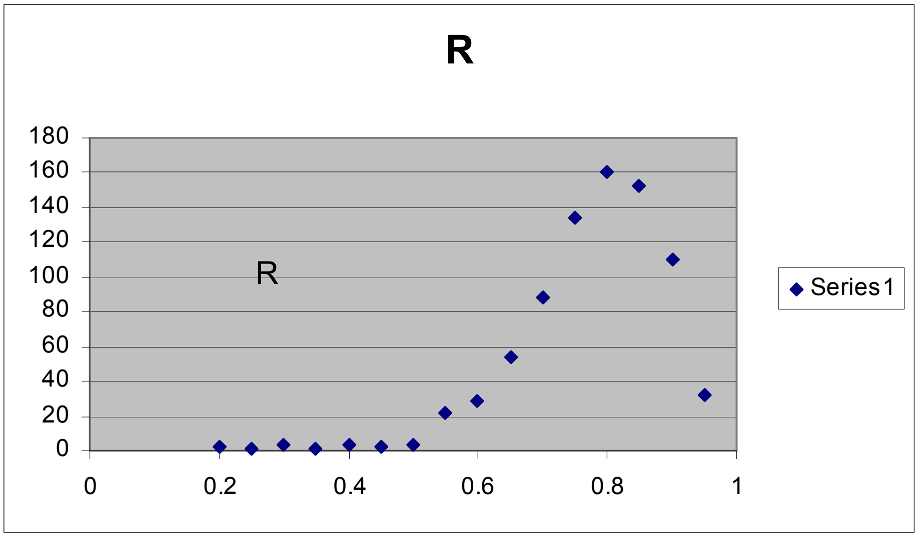


Figure 2. Box and Whisker plot of Vacha-Haase method

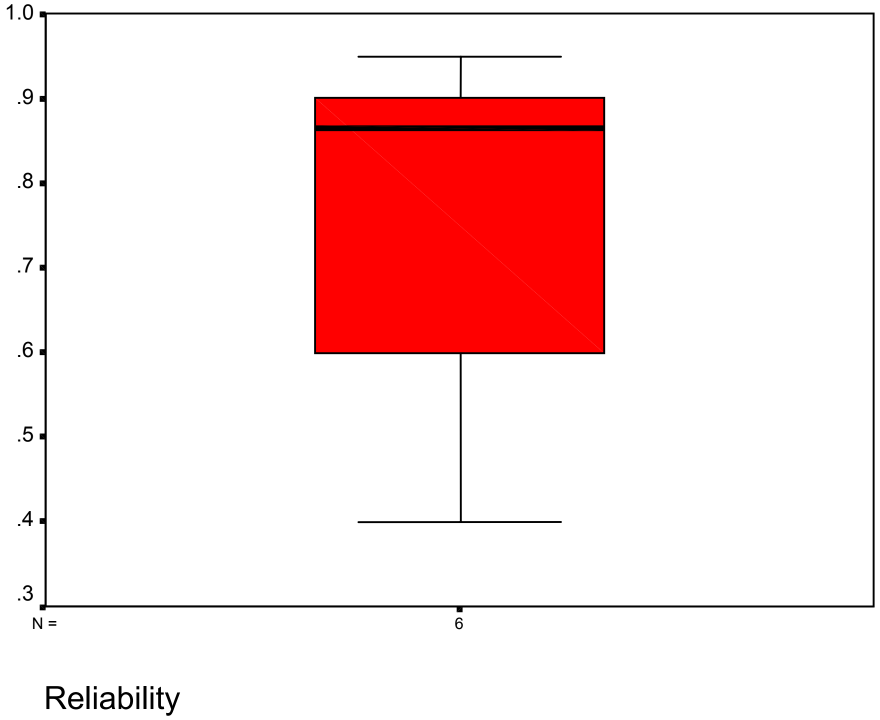


Figure 3. Representation of the moderator relationship between reliability and time

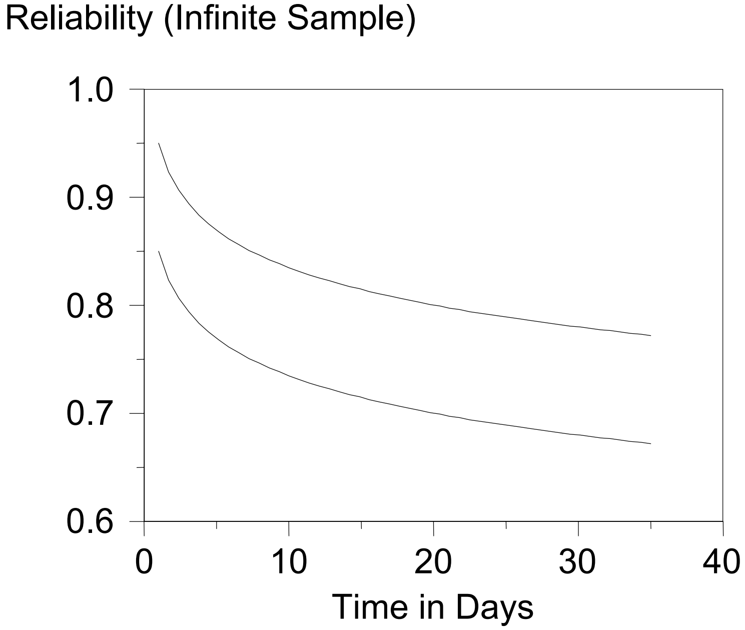
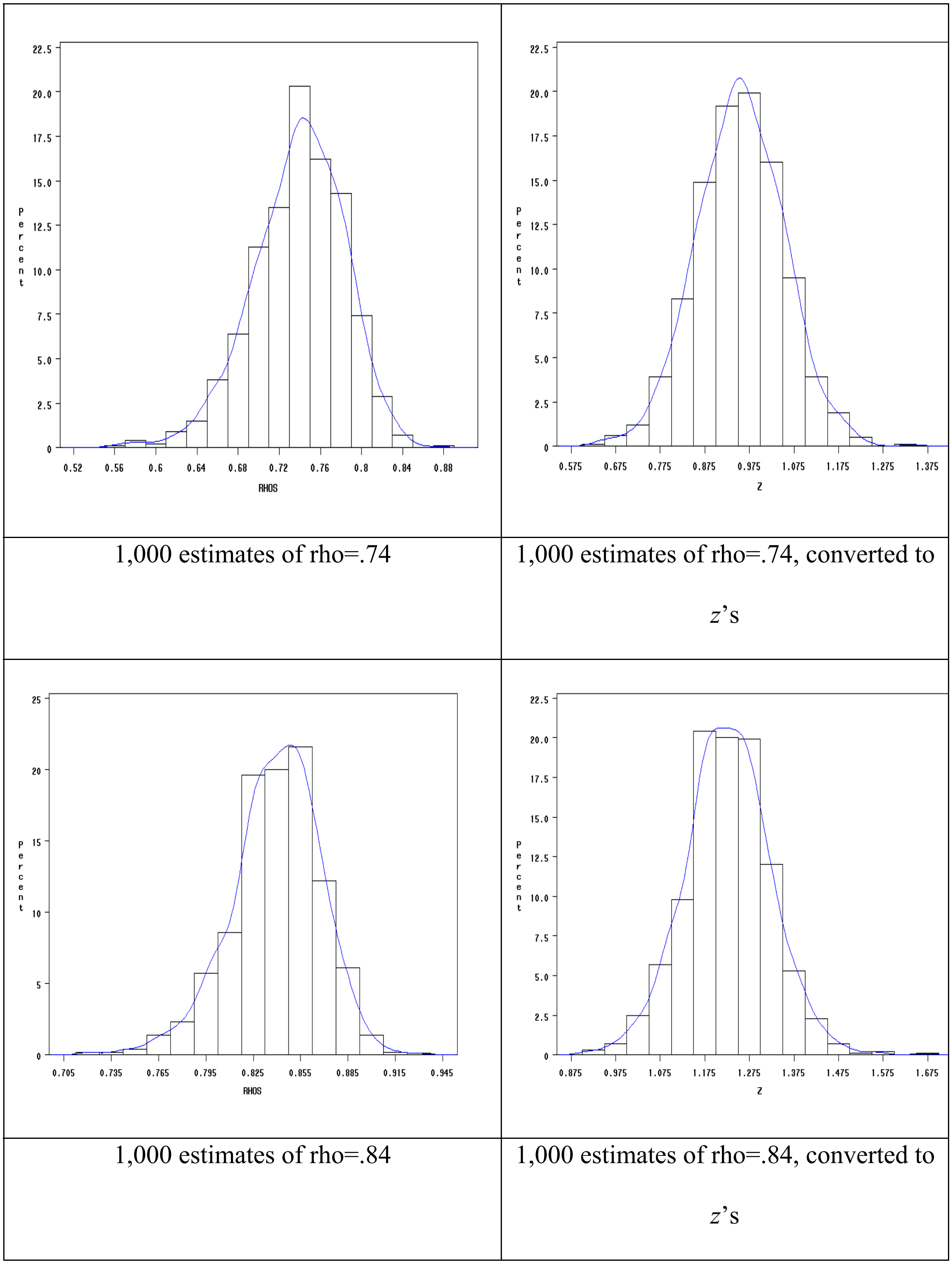


Figure 4. Sampling distributions of r's and z's no moderator conditions



1,000 estimates of rho=.74

1,000 estimates of rho=.74, converted to *z*'s

1,000 estimates of rho=.84

1,000 estimates of rho=.84, converted to *z*'s

Figure 5. Sampling distributions of r's and z's with the moderator and random error

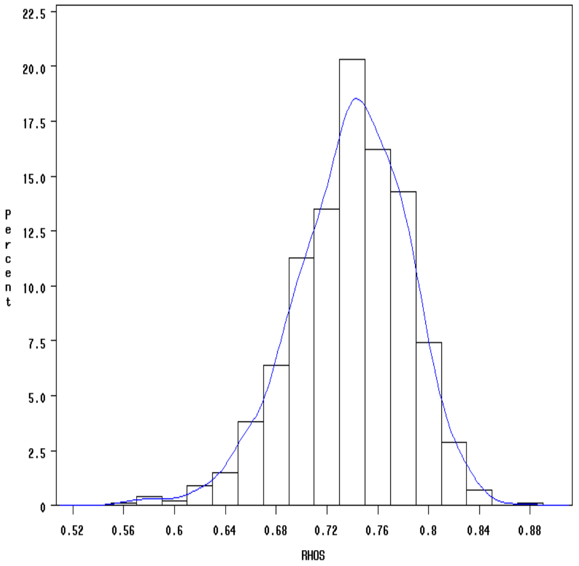
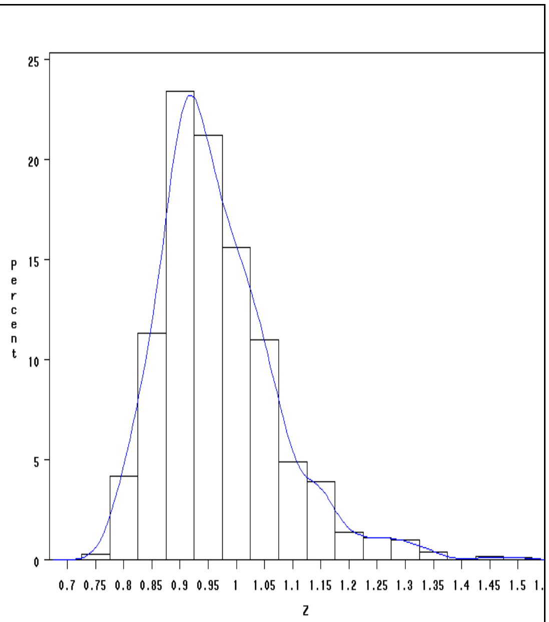
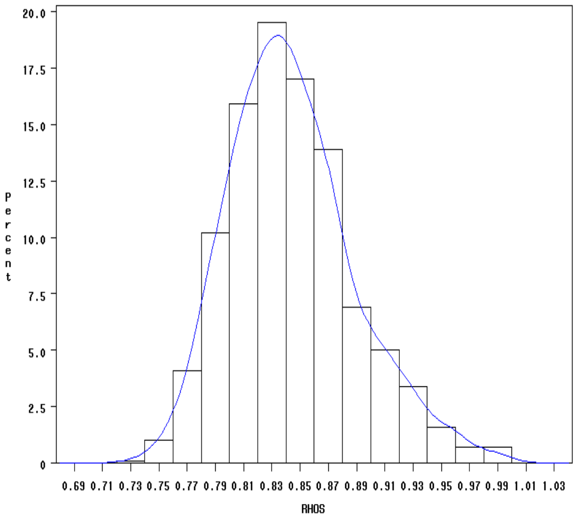
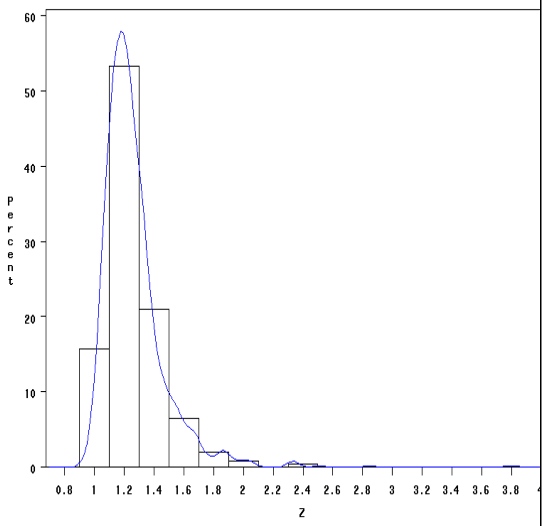
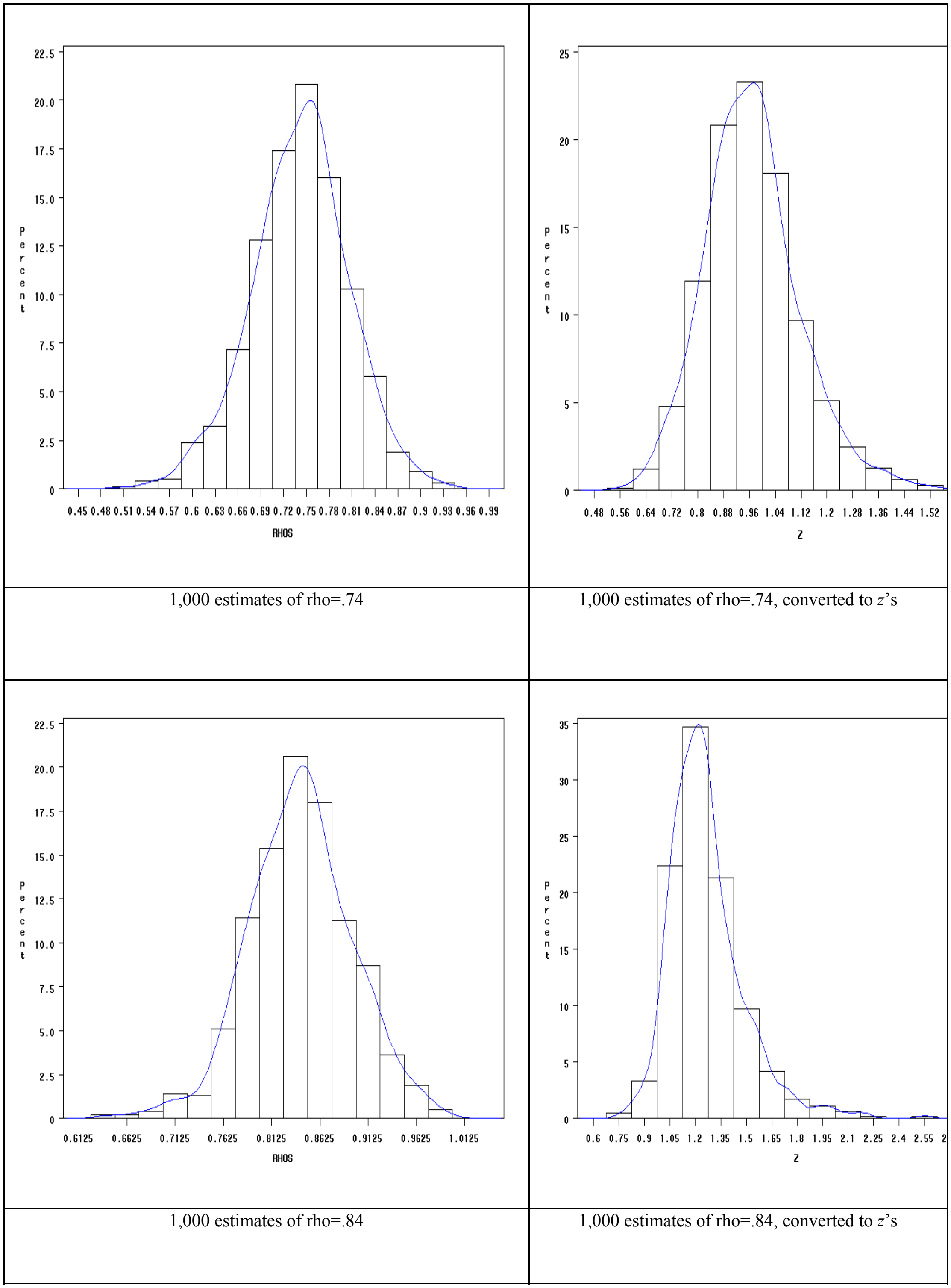
	
1,000 estimates of rho=.74	1,000 estimates of rho=.74, converted to *z*'s
	
1,000 estimates of rho=.84	1,000 estimates of rho=.84, converted to *z*'s

Figure 6. Sampling distributions of r's and z's with moderator and error but no sampling



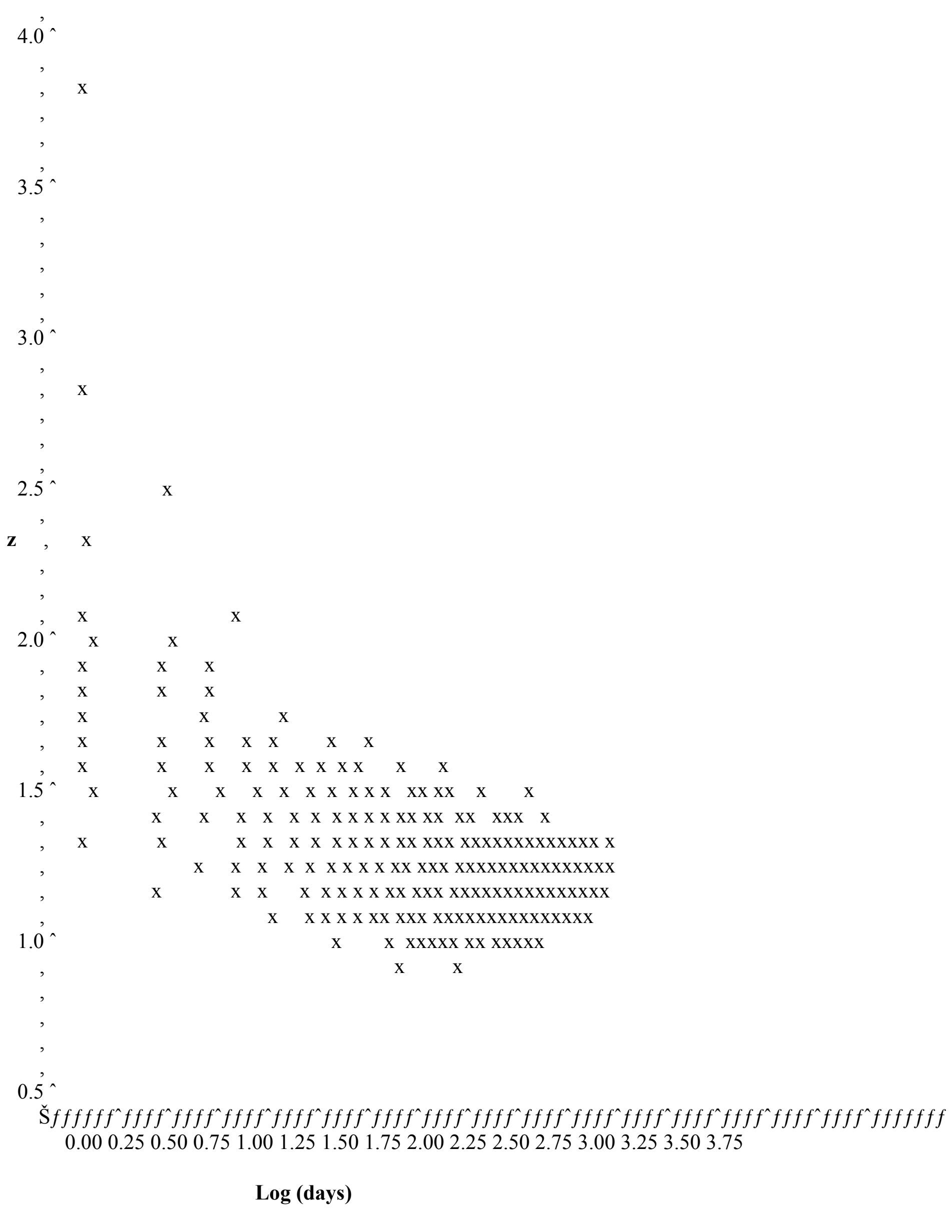
1,000 estimates of rho=.74

1,000 estimates of rho=.74, converted to *z*'s

1,000 estimates of rho=.84

1,000 estimates of rho=.84, converted to *z*'s

Figure 7. Curvilinear relationship between z’s and ln(days)



About the Author

Corinne Mason grew up in the Washington D.C. area and learned about politics, people and work from a very early age.

She attended Virginia Commonwealth University and graduated Magna Cum Laude in Psychology. From there she entered the University of South Florida's Industrial and Organizational Psychology doctoral program.

Throughout her graduate career, she also worked with several Fortune 500 companies to enhance the experience of people at work. She is currently serving as a Senior Consultant with the new Homeland Security Agency, helping the country select the correct individuals to provide national security after the September 11th terrorist attacks.

Corinne has a passion for helping people discover meaning and fulfillment in the forty plus hours that everyone one of us devotes to work each week. She is in the process of writing a book on finding motivation and meaning in work and lectures on this topic through out the U.S. She has a special focus on helping women and young people find direction and motivation in their career and work choices for life.

Corinne credits her time at U.S.F. and her mentor Dr. Mike Brannick with finding her own career direction and motivation as an Industrial and Organizational Psychologist and life coach.