2013

REVIEW AND EVALUATION OF RELIABILITY GENERALIZATION RESEARCH

Alexandra Marie Henchy

University of Kentucky, alexandra_marie@hotmail.com

Recommended Citation

Henchy, Alexandra Marie, "REVIEW AND EVALUATION OF RELIABILITY GENERALIZATION RESEARCH" (2013). Theses and Dissertations--Educational, School, and Counseling Psychology. 5.
https://uknowledge.uky.edu/edp_etds/5

This Doctoral Dissertation is brought to you for free and open access by the Educational, School, and Counseling Psychology at UKnowledge. It has been accepted for inclusion in Theses and Dissertations--Educational, School, and Counseling Psychology by an authorized administrator of UKnowledge. For more information, please contact UKnowledge@lsv.uky.edu.
STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained and attached hereto needed written permission statements(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine).

I hereby grant to The University of Kentucky and its agents the non-exclusive license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless a preapproved embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student’s advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student’s dissertation including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Alexandra Marie Henchy, Student

Dr. Fred Danner, Major Professor

Dr. Rory Remer, Director of Graduate Studies
REVIEW AND EVALUATION OF RELIABILITY GENERALIZATION RESEARCH

DISSERTATION

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the College of Education at the University of Kentucky

By
Alexandra Marie Henchy

Lexington, Kentucky

Director: Dr. Fred Danner, Professor of Educational Psychology

Lexington, Kentucky

2012

Copyright © Alexandra Marie Henchy 2012
ABSTRACT OF DISSERTATION

REVIEW AND EVALUATION OF RELIABILITY GENERALIZATION RESEARCH

Reliability Generalization (RG) is a meta-analytic method that examines the sources of measurement error variance for scores for multiple studies that use a certain instrument or group of instruments that measure the same construct (Vacha-Haase, Henson, & Caruso, 2002). Researchers have been conducting RG studies for over 10 years since it was first discussed by Vacha-Haase (1998). Henson and Thompson (2002) noted that, as RG is not a monolithic technique; researchers can conduct RG studies in a variety of ways and include diverse variables in their analyses. Differing recommendations exist in regards to how researchers should retrieve, code, and analyze information when conducting RG studies and these differences can affect the conclusions drawn from meta-analytic studies (Schmidt, Oh, & Hayes, 2009) like RG. The present study is the first comprehensive review of both current RG practices and RG recommendations. Based upon the prior research findings of other meta-analytic review papers (e.g., Dieckmann, Malle, & Bodner 2009), the overarching hypothesis was that there would be differences between current RG practices and best practice recommendations made for RG studies.

Data consisted of 64 applied RG studies and recommendation papers, book chapters, and unpublished papers/conference papers. The characteristics that were examined included how RG researchers: (a) collected studies, (b) organized studies, (c) coded studies, (d) analyzed their data, and (e) reported their results.

The results showed that although applied RG researchers followed some of the recommendations (e.g., RG researchers examined sample characteristics that influenced reliability estimates), there were some recommendations that RG researchers did not follow (e.g., the majority of researchers did not conduct an a priori power analysis). The results can draw RG researchers’ attentions to areas where there is a disconnect between practice and recommendations as well as provide a benchmark for assessing future improvement in RG implementation.
KEYWORDS: Reliability Generalization, Recommendations, Meta-analysis, Measurement, Classical Test Theory

Alexandra M. Henchy

___________________________________
Student’s Signature

04/22/2013

___________________________________
Date
REVIEW AND EVALUATION OF RELIABILITY GENERALIZATION RESEARCH

By

Alexandra Marie Henchy

Fred Danner

Director of Dissertation

Rory Remer

Director of Graduate Studies

04/22/2013

Date
I dedicate my dissertation to my cousin Lauren Savoy Olinde who completed her doctorate in pharmacy while battling cancer. She was an amazing woman who is an inspiration for me and for many others.
ACKNOWLEDGMENTS

I would like to thank God for giving the strength and perseverance to complete my dissertation. I would like to acknowledge my family for all their constant support and encouragement throughout my journey to complete graduate school.

In the academic world, I would like to thank my advisor, Fred Danner for all of his help and support from the first day I began graduate school. I would like to thank Michael Toland for being my project supervisor and working with me to turn a class project into a dissertation. I would like to thank my other committee members, Kenneth Tyler and Christia Spears Brown for helping me complete my dissertation. I would like to thank Patrick Yang for being the outside reviewer for my dissertation. I would also like to thank Angela Tombari and Allie Minieri for helping me to code all of the studies and for all of their help over the years on this project.
# Table of Contents

Acknowledgments ........................................................................................................ iii  

List of Figures ................................................................................................................ vii  

Chapter One: Introduction ............................................................................................ 1  
  Reliability .................................................................................................................. 1  
  Research Questions and Hypotheses ........................................................................ 5  
  Implications .............................................................................................................. 8  

Chapter Two: Literature Review .................................................................................. 10  
  Reliability ................................................................................................................ 10  
    Reliability estimates .......................................................................................... 11  
    Factors that influence reliability ........................................................................ 15  
  RG ......................................................................................................................... 16  
  RG Recommendations ............................................................................................. 18  
    Collecting, organizing, and coding data .............................................................. 18  
      Choosing an instrument ................................................................................. 18  
      Identifying studies .......................................................................................... 19  
      Missing data .................................................................................................... 19  
      Criteria for including studies ....................................................................... 23  
      Coding ............................................................................................................. 23  
    Analyzing and reporting results ......................................................................... 24  
      Independence of reliability reports ............................................................... 24  
      Weighting ........................................................................................................ 27  
      Homogeneity of variance .............................................................................. 28  
      Power analysis ............................................................................................... 29  
      Data transformation ....................................................................................... 30  
      Data analysis .................................................................................................. 31  
      Characterizing variability .............................................................................. 33  
  Synthesis of Recommendations .............................................................................. 34  
  Limitations of RG Research ................................................................................... 37  
  Summary ................................................................................................................. 37  

Chapter Three: Method ............................................................................................... 38  
  Sample .................................................................................................................... 38  
  Procedure ............................................................................................................... 40  
  Interrater Reliability ............................................................................................... 41  
  Data Analysis ......................................................................................................... 42  
    Data analysis of essential recommendations .................................................. 43  
    Data analysis of optimal recommendations ..................................................... 46  
    Data analysis of controversial recommendations ............................................ 47  
    Data analysis of other RG practices .................................................................. 48  
  Summary ................................................................................................................. 48
List of Figures

Figure 5.1, Percentage of studies that followed essential recommendations .......... 65
Figure 5.2, Percentage of studies that followed optimal recommendations .......... 66
Figure 5.3, Percentage of studies over time that conducted a test for homogeneity of variance ................................................................. 67
Figure 5.4, Percentage of studies that engaged in controversial practices .......... 67
Figure 5.5, Percentage of studies over time that used a weighting approach .......... 68
Figure 5.6, Number of primary studies that reported reliability estimates .......... 68
Figure 5.7, Number of reliability estimates used in the RG data analyses .......... 69
Chapter One: Introduction

According to Vacha-Haase (1998) in her seminal article that introduced the Reliability Generalization (RG) technique, RG “characterizes (a) the typical reliability of scores for a given test across studies, (b) the amount of variability in reliability coefficients for given measures, and (c) the sources of variability in reliability coefficients across studies” (p. 6). RG is a meta-analytic method that examines the sources of measurement error (i.e., factors that add imprecision into the measurement of a variable, Goodwin, 2005) for scores for multiple studies that use a certain instrument or group of instruments that measure the same construct (Vacha-Haase, Henson & Caruso, 2002). For example, Yin and Fan (2000) conducted a RG study using the Beck Depression Inventory (BDI), and the reliability estimates for the BDI scores were summarized across studies. Throughout the present paper, the term instrument is used for the purpose of simplicity, but it is important to note that the words test and measure would also be appropriate terms.

Henson and Thompson (2002) noted that, as RG is not a monolithic technique; researchers can conduct RG studies in a variety of ways and include diverse variables in their analyses. Although many recommendations have been proposed for conducting RG studies, differing recommendations exist in regards to how researchers should retrieve, code, and analyze information when conducting RG studies.

Reliability

Reliability can be examined from the perspective of classical test theory (CTT). CTT assumes that every participant has a “true” score that would be attained if there were not any errors in measurement. However, instruments that are created by researchers are
not perfect, thus the observed score of participants usually differs from the “true” score (Kaplan & Saccuzzo, 2005). In a conceptual sense, an observed score of a participant has two parts: one part is someone’s “true” score and the other part is an “error” score, which is attributable to the inaccuracy of measurement. If scores have a greater proportion of error, then reliability is lower; however, if scores have lower amounts of error, the reliability is higher (Weirsma & Jurs, 2009). Dimitrov (2002) stated that in empirical research, “true” scores cannot be directly ascertained, and so the reliability is usually assessed by reliability estimates.

Authors of individual studies have been encouraged to report the reliability estimates for the instruments they use based on the scores obtained from their own samples. Wilkinson and the APA Task Force on Statistical Inference (1999) stated that “a test is not reliable or unreliable . . . thus, authors should provide reliability coefficients of the scores for the data being analyzed even when the focus of their research is not psychometric” (p. 5). Additionally, sample reliability estimates vary across studies, necessitating researchers to evaluate and report the reliability of the scores from their data rather than rely on prior reliability estimates (Romano & Kromrey, 2002).

Knowledge of sample reliability estimates offers valuable information as statistical analyses and interpretations of scores in a primary study are contingent upon reliability evidence (Warne, 2008). An added benefit of reporting reliability estimates for instrument scores derived from study samples is the meta-analytic opportunities offered for researchers who conduct RG studies. If authors of primary studies (i.e., studies conducted by researchers who use the instrument[s] of interest) do not report the
reliability estimates of their samples, it is challenging for researchers who conduct RG studies.

RG can be used to examine the reliability of instrument scores and is a technique that can be employed to help researchers gain a better understanding of reliability (Warne, 2008). One problem in the current literature is that many researchers misunderstand the concept of score reliability (Thompson, 1999; Vacha-Haase, 1998). Warne (2008) argued that many researchers do not recognize that reliability estimates originate in the sample data they gather and not in instruments. This misunderstanding has led authors to make statements such as “the test is reliable” (Vacha-Haase, 1998, p. 6). However, reliability is a property of scores not instruments (Crocker & Algina, 1986).

The belief that reliability is a property of instruments is problematic as it has led to an underreporting of reliability estimates and a common dismissal of its importance in research (Cousin & Henson, 2000). In their review of RG practices, Vacha-Haase and Thompson (2011) examined the reliability reporting practices of RG studies from 1998 through 2010. The results of their analyses showed that the majority of authors of primary studies did not mention reliability or report the reliability of their own scores. The results of the study by Vacha Haase and Thompson also showed that some authors engaged in the practice of reliability induction. According to Vacha-Haase, Kogan and Thompson (2000), reliability induction is the practice of reporting reliability estimates from prior samples (or even manuals) and not the reliability estimates of one’s own scores. Reliability induction is problematic because of the potential negative effects of low score reliabilities on subsequent data analyses (Thompson & Vacha Haase, 2000).
Vacha-Haase et al. (2000) argue that if researchers want to use reliability induction, they need to examine whether their sample and the prior sample are comparable in terms of composition and score variability. However, since score reliability varies from one study (or sample) to another it should be examined in every study (Thompson, 1999), and reliability induction is not recommended as standard practice.

Reliability is important for a variety of reasons. For example, Warne (2008) noted that reliability is a necessary, but not sufficient condition for validity (which is an evaluative review of both the evidence used for and the results of score interpretation and use, Messick, 1995). Therefore, a low score reliability can influence the validity of an instrument’s use. Furthermore, unreliability of instrument scores can be a threat to statistical conclusion validity (which is defined as the validity of inferences regarding the correlation between the predictor variables and outcomes, Shadish, Cook, & Campbell, 2002).

Reliability is also important because conducting analyses with scores in the outcome variable with low reliability weakens statistical power against Type II errors (Warne, 2008). A Type II error or false negative occurs when researchers fail to reject a wrong null hypothesis (Wiersma & Jurs, 2009). Furthermore, effect size can be limited by score reliability (Helms, 1999). In general, effect size examines/measures the amount of influence that one variable has on another variable (Goodwin, 2005). For example, Reinhardt (1996) noted that if a dependent variable is measured in a way that scores are entirely unreliable, the effect size in the study will inevitably be zero; the results of the study will not be statistically significant even if researchers have a very large sample size.
Research Questions and Hypotheses

The purpose of the present study was to answer three research questions: (a) What recommendations have been made for conducting RG studies? (b) What are the current practices of researchers conducting RG studies? (c) How do the current practices of RG researchers compare to the RG recommendations? It is important to address these questions because it provides RG researchers with a better understanding of the RG literature, examines the current practices of RG researchers, and can be used to improve future RG research. For the purpose of the present paper, applied RG studies refers to any study that conducted an RG analysis, and recommendation studies are those that provide suggestions about how best to conduct RG studies.

Since RG studies are a type of meta-analysis, one can look to the meta-analytic literature for recommendations concerning best practices and for areas of concern where best practices are often not followed. In the present study, best practices are those which many researchers support and for which there is not great controversy regarding the practice. For example, Dieckmann, Malle, and Bodner (2009) conducted a review of meta-analytic research and found there to be a disconnect between recommendations for conducting meta-analytic studies and the common practices of researchers. The present study is the first comprehensive review of both current RG practices and RG recommendations. Based upon the prior research findings of other meta-analytic review papers (e.g., Ahn, Ames, & Myers, 2012; Dieckmann et al., 2009), the overarching hypothesis was that there would be differences between current RG practices and best practice recommendations made for RG studies. Publication bias and statistical power analysis are used in the present section to illustrate the hypothesis that there are
differences between recommendations for and common RG practices. Publication bias is the bias that occurs when studies are more likely to be published when their results are statistically significant (Begg, 1994), and a power analysis is a computation of statistical power, which allows researchers to know whether investigations of hypotheses are likely to detect the anticipated effects (Hedges & Pigott, 2001).

According to Howell and Shields (2008), many RG researchers are concerned with the influence that unpublished studies may have on the results of their study; this issue is also faced by other meta-analytic researchers (Howell & Shields, 2008). Ahn et al. (2012), Dieckmann et al. (2009), and Geyskens, Krishnan, Steenkamp, and Cunha (2009) examined publication bias in their reviews of meta-analytic studies and reported that the bulk of the meta-analyses examined in their studies did not examine publication bias. Therefore, it was hypothesized that a similar disconnect would exist in the RG literature between the recommended practice of examining and reporting of any publication bias by RG researchers.

Another area where there may be a disconnect between RG recommendations and applied RG papers is in the practice of conducting power analyses. Dieckmann et al. (2009) examined power analyses in their meta-analytic review. They noted that low statistical power can influence the statistical conclusion validity of the results of meta-analytic studies. Therefore, Dieckmann et al. recommended that power analyses be conducted and reported. Similarly, Hedges and Pigott (2001) noted that it is important for researchers to conduct power analyses before conducting meta-analyses because researchers do not want to begin a meta-analytic study if there is only a small chance that their results will be useful. However, Dieckmann et al. found in their study that only one
meta-analytic study out of 100 studies conducted a retrospective (post-hoc) power analysis, and no studies conducted an a priori power analyses. Therefore, it was hypothesized that many applied RG studies would not conduct an a priori or post-hoc power analysis, although it is a highly recommended practice, especially when applying for funding from granting agencies.

In order to answer the aforementioned research questions and hypotheses, a search was made for all applied and recommendation RG studies using the keywords reliability, generalization, meta-analysis, and combinations of these words. The search was conducted in the PsycINFO, ERIC, and Dissertation Abstracts online databases. Data consisted of relevant RG applied and recommendation papers, book chapters, and unpublished papers/conference papers identified by the searches.

After the search for the papers, the RG recommendation papers were reviewed. Additionally, as common RG recommendations emerged from reading the papers they were also recorded and discrepancies in the recommendations of various authors were noted.

A coding scheme was created based upon the RG procedures suggested in the recommendation papers and the meta-analytic review paper by Dieckmann et al. (2009). Every applied RG paper was coded by two researchers and disagreements were resolved by reviewing and discussing the coding discrepancies for consensus. After examining the code sheet, which was used for coding, it was determined that there were five overarching characteristics included in the code scheme. The overarching characteristics that were coded included how RG researchers: (a) collected studies, (b) organized studies, (c) coded studies, (d) analyzed their data, and (e) reported their results.
As noted by Dieckmann et al. (2009) when reviewing studies, it is often hard to differentiate between what was practiced by researchers and what was reported. Although overall trends in the results of the RG studies were reported, it is possible that researchers did engage in some practices (such as conducting an a priori power analysis) but did not report it in their studies due to various reasons (e.g., space restrictions, editorial decisions to remove unnecessary content). However, it is only possible to analyze RG practices that were described in the studies. Therefore, one limitation of the present study was that the coding only included what was reported in RG studies even though it is possible that researchers conducted practices they did not report in their studies. Additionally, editors or reviewers may have asked authors to perform certain analyses or include information that is frequently found in other studies, but which was not suggested by recommendation papers.

**Implications**

RG is an important meta-analytic method because RG can alter researchers’ thinking about reliability issues (Henson & Thompson, 2002). Additionally, Vacha-Haase et al. (2002) argued that the results of RG studies can have important implications for researchers who want to acquire a theoretical understanding of reliability (e.g., the influence of reliability or measurement error on effect size).

It is important to conduct RG studies because RG findings promote understanding of factors that influence score reliability and guide researchers in making multi-item instruments (e.g., scales, questionnaires) that produce more reliable scores (Cousin & Henson, 2000). Warne (2008) noted that the knowledge of the instrument forms or circumstances under which an instrument produces high score reliability within a sample
will improve researchers’ decisions when conducting studies. Vacha-Haase et al. (2002) noted that RG studies aid researchers who are interested in using an instrument to help in making decisions both in individual cases and within groups.

Studying the differences in RG practices by researchers is important as procedural decisions such as how researchers retrieve, code, and analyze information can affect the conclusions drawn from meta-analytic studies (Schmidt, Oh, & Hayes, 2009), such as RG. Although this particular type of meta-analytic method is relatively new, meta-analysis is not, and enough time has passed to evaluate the current state of the RG literature.

Many researchers spend both time and effort writing recommendation papers for applied researchers. However, researchers do not always follow recommendation papers (see e.g., Dieckmann et al., 2009). RG researchers should be aware of RG recommendations, but if the researcher has a valid reason for not adhering to a recommended practice, he or she should inform the reader why he or she deviated from the recommended practice.

This is the first research project, to our knowledge, that provides both a comprehensive examination of RG recommendations and assesses the progress of RG practices since its inception. The results of this study provide a comprehensive guide to the sometimes controversial RG recommendations, which may be confusing to some practitioners. The results of this study can also offer a guide for conducting RG studies as well as a benchmark for assessing future improvement in RG implementation.
Chapter Two: Literature Review

RG is a method that can be used to describe and investigate variance in score reliability (Vacha-Haase, 1998). Cousin and Henson (2000) argued that RG highlights the variation in score reliability that can transpire across studies using the same instrument. RG helps researchers to understand that “reliability is not an immutable unchanging property of tests stamped indelibly into booklets during the printing process” (Henson & Thompson, 2002, p. 124).

Warne (2008) argued that reliability is not one single characteristic of an instrument as there are multiple potential sources of measurement error and different methods to measure them. As instruments are not inherently reliable, score reliability varies from administration to administration, and therefore should be evaluated in all studies (Helms, 1999). The following literature review examines reliability, RG, and recommendations for RG practices.

Reliability

According to the standards written by the American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME; 1999) reliability refers to the consistency of measurements when a testing process is repeated for an individual or group of individuals. Essentially, reliability refers to the consistency of measurements (Cohen, Swerdlik, & Phillips, 1996; Wiersma & Jurs, 2009). According to Sawilowsky (2000b) there are a variety of definitions for reliability. Some definitions of reliability are theoretical, for example, the proportion of true score variance to total score variance (variance is defined as the differences between scores for the participants in a study,
Warner, 2008). Sawilowsky (2000b) noted that other definitions of reliability describe the procedures by which reliability evidence is attained. For example, reliability could be indexed by the correlation of the scores on an instrument with the scores the following time or occasion the instrument is administered. However, the description of the process of how reliability evidence is attained does not provide a definition of reliability (Sawilowsky, 2000b).

According to Cortina (1993), one vital detail to note is that the level of score reliability that is acceptable for a study varies. When finer distinctions between scores have to be made, the acceptable level of reliability must also be higher (e.g., needing to distinguish on an achievement test between a score of 800 and 400 versus 759 and 760, Cortina, 1993). Additionally, Kane (2011) noted that the tolerance for error depends on possible problems or adverse outcomes produced by measurement error (e.g., misclassification). Kane defined the tolerance for error as the magnitude at which the errors in a certain context start to interfere with the planned interpretations and uses of instrument scores.

**Reliability estimates.** In general, there are four broad types of reliability: test-retest reliability, parallel forms reliability, internal consistency of reliability, and interrater reliability (Kaplan & Saccuzzo, 2005). As previously discussed in the introduction section, Dimitrov (2002) stated that in empirical research, “true” scores cannot be directly ascertained, and so the reliability is usually estimated by reliability estimates.

Test-retest is one type of reliability estimate. The estimate of test-retest reliability is also known as the coefficient of stability (Cohen et al., 1996). Kaplan and Saccuzzo
(2005) noted that test-retest reliability estimates evaluate the reliability of instrument scores when an instrument is given at multiple and subsequent points in time. According to Cohen et al., (1996) test-retest is an estimate of reliability attained by correlating pairs of scores from the same people on multiple instrument administrations. Dimitrov (2002) contended that test-retest reliability estimates are most appropriate for evaluating traits that are stable across the time period the instrument is given (e.g., work values or personality). According to Kaplan and Saccuzzo when administering instruments at multiple points in time, researchers should be aware of potential carry-over effects [i.e., when the first time an instrument is administered influences the subsequent time(s) the instrument is administered]. Therefore, the time intervals chosen for administering an instrument should be carefully selected (Kaplan & Saccuzzo, 2005).

The parallel forms procedure for estimating reliability entails the use of two or more equivalent forms of an instrument, which is given to individuals with a short time between administrations (Wiersma, & Jurs, 2009). According to Cohen et al. (1996), the terms parallel forms and alternate forms are sometimes used interchangeably in the literature. However, parallel forms are versions of an instrument built from the same test specifications, which have equal means and variances of the observed scores, but different items sampled from the same broad domain being measured. Alternate forms are simply different versions of the same instrument (Cohen et al., 1996). According to Kaplan and Saccuzzo (2005), parallel forms reliability estimates evaluate scores from multiple forms of an instrument that measure the same construct. The various forms of the parallel instruments use diverse items; however, the item difficulties for the different forms should be the same in content. It is possible that the parallel forms are
administered to people on the same day or on different days. However, in practice, researchers may find it difficult to develop multiple forms of an instrument and therefore may only create one form (Kaplan & Saccuzzo, 2005).

Internal consistency reliability estimates examine the reliability of instrument scores within one particular instrument (Kaplan & Saccuzzo, 2005). The internal consistency reliability estimate refers to the intercorrelations between items on the same instrument (Kaplan & Saccuzzo, 2005). There are multiple types of internal consistency reliability estimates such as KR20. Kuder and Richardson (1937) provided the KR20 formula, which is used to calculate the reliability of instrument scores with items that have a dichotomous format. Kuder and Richardson defined KR20 as follows:

\[ r_{tt} = \frac{n}{n-1} \times \frac{\sigma_t^2 - n\bar{pq}}{\sigma_t^2}, \]

where \( r_{tt} \) is the score reliability from an instrument; \( n \) is the number of items in the instrument; \( \sigma_t^2 \) is the obtained instrument variance; \( n\bar{pq} \) is the variance of \( n \) equally difficult items when they are uncorrelated. Kuder and Richardson’s technique considers all potential ways of splitting the items on an instrument when estimating score reliability.

Cronbach’s coefficient alpha is a more general internal consistency reliability estimate when compared to KR20 (Cronbach, 1951) because it can be used with instruments that do not have a dichotomous format (Kaplan & Saccuzzo, 2005). Cronbach defined alpha as follows:

\[ \alpha = \frac{n}{n-1} \left(1 - \frac{\sum V_i}{V_t}\right), \]
where $\alpha$ is the estimate of reliability; $n$ is the number of items in the instrument; $\sum_i$ is the sum of the items; $V_i$ is the variance of item scores after weighting, and $V_t$ is the variance of instrument scores. Cronbach’s coefficient alpha is one of the most frequently used ways of estimating internal consistency of reliability (Dimitrov, 2002). According to Reinhardt (1991) alpha is a lower bound estimate of score reliability, or in other words a conservative estimate of reliability. Additionally, Cortina (1993) argued that an acceptable coefficient alpha value suggests only that, on the average, the split halves of the instrument are highly correlated. The coefficient alpha value does not determine the extent to which the split halves are measuring the construct of interest.

Interrater reliability estimates examine the agreement of different raters who evaluate the same variables (Kaplan & Saccuzzo, 2005). Kaplan and Saccuzzo (2005) noted the most common way to calculate interrater reliability is to note the percentage of times that two or more raters agree. However, interrater reliability can also be calculated by use of the kappa statistic, which measures the agreement between two judges (Cohen, 1960). According to Cohen (1960), the coefficient $k$ is the proportion of agreement after chance agreement is taken out of consideration. Cohen defined coefficient $k$ as follows:

$$k = \frac{p_o - p_c}{1 - p_c}$$

where $p_o$ is the proportion of units in which the raters agree, and $p_c$ is the proportion of units for which agreement expected by chance.

Each reliability type examines a different source of measurement error. When researchers want to examine score reliability, they ought to determine the source of measurement error they would like to evaluate (Kaplan & Saccuzzo, 2005). Cortina (1993) contended that if error factors that are related to time are of interest, than test-
retest or parallel forms reliability may be used. If error factors that are related to different items on the same instrument are of interest, then internal consistency reliability estimates could be used. If errors related to differences among raters are of interest, then interrater reliability may be used.

**Factors that influence reliability.** It is important to be aware of different factors that may influence reliability. Dimitrov (2002) argued that researchers should recognize and discuss how these factors may limit their procedures and results. Both instrument characteristics and sample characteristics have been used in RG studies to examine whether the characteristics were predictors of variability in score reliability estimates. In Vacha-Haase and Thompson’s (2011) review of RG research, they found that the most commonly used predictor variables were gender, sample size, age in years, and ethnicity. The results of their review showed that the four predictors that researchers used that were typically notable (i.e., better predictors of the variabilities in score reliabilities) were instrument length, the score standard deviation in the primary studies (i.e., individual studies examined in the RG study), participant age, and participant gender.

Helms (1999) stated that how homogeneous or heterogeneous a group of participants is influences the total instrument score variance. For example, if an instrument is given to a group of graduate students in the same program with the same background and grade point average they will probably answer the questions in a similar manner, thus reducing the variability in the overall instrument scores and decreasing coefficient alpha (Helms, 1999).

Kaplan and Saccuzzo (2005) noted that internal consistency reliability estimates are influenced by instrument length, with the reliability of the scores increasing as
instrument length increases; however, Warne (2008) argued that this is not always true. In order to investigate the influence of length on coefficient alpha, Cortina (1993) calculated coefficient alpha for scales with different numbers of items and different average item intercorrelations. The results of the calculations by Cortina (which examined alphas for a variety of conditions) showed that the number of items had a strong influence on coefficient alpha, particularly when there were low levels of average item intercorrelations.

Some factors can also influence test-retest reliability. For example, Kaplan and Saccuzzo (2005) contended that test-retest reliability is influenced by the amount of time that passes between administrations. If two administrations of an instrument are given close together in time, it is possible there is a greater risk of carryover effects due to practice. Carryover effects are reduced when there is more time between administrations. However, longer intervals between administrations often results in low test-retest reliability estimates.

**RG**

RG is an extension of validity generalization (Schmidt & Hunter, 1977). In validity generalization studies, features of the primary studies (e.g., sample size) are investigated to determine which characteristics influence the variations in the validity coefficients. RG is an important method because researchers can determine which factors may lead to higher reliability estimates by examining the samples, instrument forms, or circumstances under which an instrument is taken (Warne, 2008).

Deditius-Island and Caruso (2002) noted that RG is a meta-analytic technique that examines the reliability of scores of a particular instrument in a much broader way than
could be attained by any one study. According to Romano and Kromrey (2002), meta-analysis is a quantitative research design that can review large bodies of literature. Meta-analyses change individual study results to a common metric and evaluate them across studies. Additionally, Matt and Cook (2009) noted that at the center of every research synthesis is a relationship researchers want to learn something about which cannot be determined with confidence from a single study.

According to Romano and Kromrey (2002), researchers who conduct RG studies endeavor to characterize the psychometric properties of a hypothetical universe of studies that may use a certain instrument. The psychometric properties may consist of the research design features that may influence the reliability estimates and the variance of the reliability estimates across studies. RG can be used to examine the reliability estimates of one instrument or different instruments that measure the same construct (Romano & Kromrey, 2002).

Cousin and Henson (2000) noted that in a RG study, the primary studies are the unit of analysis and the reliability estimates are the dependent variables. The independent variables in RG studies are the instrument and study characteristics selected that may influence the variation in reliability estimates. The results of RG studies can give information about different sources (e.g., instrument length) that may be producing measurement error across studies that use a certain instrument. This information helps researchers determine what factors influence score reliability (Cousin & Henson, 2000).

RG studies can contribute to the research literature in multiple ways. For example, RG can alter researchers’ thinking about reliability issues (Henson & Thompson, 2002). For instance, Warne (2008) affirmed that RG research can have
results that go against intuition (e.g., longer instruments are not always more reliable, Kieffer & Reese, 2002). Counterintuitive findings may lead researchers to re-examine their assumptions about reliability. Additionally, Warne noted that RG results help practitioners improve their understanding of the instruments that they use. If practitioners are aware of the populations an instrument is appropriate for or which subscale(s) have low reliability estimates, they can make informed decisions when investigating instrument scores and will be aware of which instrument is suitable for a certain set of circumstances.

**RG Recommendations**

The purpose of examining RG recommendations was to determine the state of current recommendations for RG studies. The present study provides a comprehensive guide to the many and sometimes controversial RG recommendations, which may be confusing. It is important to note that RG studies are limited by the information provided in primary studies. Therefore, not all recommendations are applicable to every RG study. Additionally, researchers should be familiar with their data and be aware of how the RG recommendations could influence their procedures, data, and results.

**Collecting, organizing, and coding data.** When conducting a RG study researchers must first collect, organize, and code the primary studies that will be used as the data for their studies.

**Choosing an instrument.** Henson and Thompson (2002) noted that any instrument could be used for a RG study that has scores for which reliability can be calculated. Additionally, when choosing an instrument for a RG study, Cousin and
Henson (2000) argued that any achievement or attitudinal instrument could be used so long as enough studies exist to justify a synthesis of research.

**Identifying studies.** After selecting an instrument, RG researchers must assemble the studies that use the instrument of interest (Cousin & Henson, 2000; Thompson, 19999). Henson and Thompson (2002) stated that databases such as PsycINFO can help researchers find studies. Warne (2008) noted that prior researchers have searched for studies by searching Dissertation Abstracts (Youngstrom & Green, 2003), the references of meta-analytic studies (Li & Bagger, 2007) and contacting well-known researchers who use a certain instrument to ask for reliability data (O’Rourke, 2004). When searching for primary studies for a RG study, researchers should use keywords that are broad enough to capture the different forms of the instrument, including abbreviations (Henson & Thompson, 2001).

**Missing data.** In RG studies, there are multiple types of missing data. For example, missing data can occur in RG studies when primary studies do not report the reliability of their samples. Warne (2008) noted that a problem with RG studies is the limited number of primary studies that can be included due to low reporting rates of reliability estimates. However, RG researchers must account not only for the potential influence of studies that do not report reliability information, but also for the influence of unpublished studies (Vacha-Haase et al., 2002).

Missing data can also occur in RG studies because of unpublished papers. According to Romano and Kromrey (2002), meta-analyses are typically performed using only published studies; this is problematic as published studies may be biased towards statistically significant results. Rosenthal (1979) contended that for any known research
topic, one does not know how many studies have been performed, but never published; this is also known as the file-drawer problem. The file-drawer problem is a problem facing any researcher conducting a meta-analytic study (Romano & Kromrey, 2002). File-drawer studies are problematic because the exclusion of unpublished studies may lead to biased meta-analytic estimates if their psychometric properties (such as reliability) are different from the psychometric properties of published studies (Howell & Shields, 2008).

In order to address the problem of missing data due to the file-drawer problem, researchers can test for publication bias by using methods such as the funnel plot technique (Light & Pillemer, 1984). According to Light and Pillemer (1984), the funnel plot technique reveals potential publication bias from an underrepresentation of studies in a literature review. In a display of a funnel plot, the sample size is on the y-axis and the effect size is on the x-axis, and a dot or other marker represents each study. If there is not a publication bias, then the plot should look like an inverted funnel (i.e., there is a broad spread of dots for the highly variable smaller studies at the base and the spread decreases as the sample size increases forming a funnel that looks like a waffle cone).

Howell and Shields (2008) created the Fail-Safe N for RG to examine the influence of unpublished studies. This method or equation estimates the influence of both published papers that do not report reliability estimates and unpublished papers. Howard and Shields’ Fail-Safe N is an extension of the Fail-Safe N developed for use in traditional meta-analyses (see e.g., Orwin, 1983; Soeken & Sripusanapan, 2003). The Fail-Safe N equation estimates a realistic ‘what if’ scenario, yet worst-case, average score reliability estimate for an instrument, assuming that the nonreporting studies have much
lower average reliability estimates when compared to studies that report reliability.

According to Howell and Shields, the Fail-Safe $N$ equation for RG estimates the number of file-drawer studies needed to drop or reduce the overall score reliability below a particular criterion value. The use of the Fail-Safe $N$ formula can inform RG researchers whether their instruments’ mean score reliability estimates are reasonable portrayals of the population parameters or whether additional reliability estimates are necessary to accurately estimate the population reliability (Howell & Shields, 2008). Howell and Shields noted that when a researcher’s estimate is below the lowest acceptable reliability estimate, than the results of the RG study should be tempered until additional studies that report reliability estimates are gathered.

Howell and Shields (2008) defined the Fail-Safe $N$ as follows:

$$\text{Fail Safe } N = N_{\text{RG Sample}} \times \frac{\alpha_{\text{UW RG Sample}} - \alpha_{\text{Threshold}}}{\alpha_{\text{Threshold}} - \alpha_{\text{File-Drawer}}}$$

(4)

where $N_{\text{RG Sample}}$ is the number of studies identified as reporting reliability estimates; $\alpha_{\text{UW RG Sample}}$ is the unweighted average reliability estimate calculated in the RG; $\alpha_{\text{Threshold}}$ is the lowest acceptable score reliability or threshold of the instrument; and $\alpha_{\text{File-Drawer}}$ is file-drawer unweighted average reliability estimate computed in the RG. If researchers are interested in using weighted mean reliabilities and file-drawer studies, they can multiply the above equation by: $(\text{Weight}_{\text{RG Sample}}/\text{Weight}_{\text{File-drawer}})$. When researchers multiply the above equation by weight, they need to assign a weight to both the RG sample and the file-drawer sample. For more information regarding how to compute the weighted Fail-Safe $N$, see Howell and Shields.

In order to address the problem of missing data in RG studies due to primary studies not reporting the reliability coefficient, Cousin and Henson (2000) recommended
using the $KR_{21}$ formula. The $KR_{21}$ formula requires items to have a dichotomous format (e.g., correct vs. incorrect), the total number of participants sampled, the mean score of the dependent variable, and either the variance or the standard deviation of the instrument scores (Kuder & Richardson, 1937). Kuder and Richardson (1937) defined $KR_{21}$ as follows:

$$r_{tt} = \frac{n}{n - 1} \cdot \frac{\sigma_t^2 - npq}{\sigma_t^2},$$  \hspace{1cm} (5)$$

where $r_{tt}$ is the reliability of the sample; $n$ is the number of items in the instrument; $\sigma_t^2$ is the variance obtained from the instrument scores. Additionally, for the term $npq$, $\bar{p}$ is the mean $p$ (which is the number of people passing an item), $\bar{q}$ is the mean $q$ (which is the number of people failing an item), for the variance of $n$ equally difficult items when they are uncorrelated. One of the assumptions of the $KR_{21}$ formula is equal item difficulty (Kuder & Richardson, 1937). The $KR_{21}$ assumption of equal item difficulty is seldom met in practice; therefore, the $KR_{21}$ formula may underestimate reliability (Kaplan & Saccuzzo, 2005). Even though the $KR_{21}$ formula may underestimate reliability, it still can help researchers address the issue of missing data in RG studies that use instruments that have scales with a dichotomous format.

There are also other ways that RG researchers can address the issue of missing data when reliability is not reported in a study being used in the RG analysis. For example, the results of a review of RG studies by Vacha-Haase and Thompson (2011) found that some RG researchers contacted the authors of primary studies to ask them if reliability information was available for their studies.

Another way to address the issue of missing data is by means of multiple imputation (see e.g., White, Royston, & Wood, 2010). In Romano and Kromrey’s (2002)
Monte Carlo study, which examined missing data treatments, they found that the multiple imputation approach was better than the listwise deletion approach. They noted that the practice of listwise deletion can lead to estimates that are very inaccurate. Briefly, multiple imputation involves using the distribution of the observed data to estimate a set of reasonable values for the missing data (White et al., 2010). For readers who are interested in the topic of multiple imputation, White et al. provided a tutorial for conducting multiple imputations using chained equations.

**Criteria for including studies.** Dieckmann et al. (2009) noted that the inclusion criteria used by meta-analytic researchers for studies included in the meta-analysis should be specified. Additionally, Wilkinson and the APA Task Force on Statistical Inference (1999) stated researchers should define their populations (which can include participants or studies) because the interpretations of the results depend on the features of the population. As previously noted in the literature review section, meta-analyses are typically performed using only published studies (Romano & Kromrey, 2002). RG researchers need to consider whether to include journal articles, book chapters, dissertations, and unpublished papers presented at conferences in their data analyses.

**Coding.** One of the goals of a RG study is to determine what characteristics of the sample and instrument influence score reliability. In Vacha-Haase and Thompson’s (2011) review of RG studies, they found that RG researchers typically coded features from the primary studies that might predict variability in the score reliabilities (e.g., gender and sample size). When creating a coding schema, researchers need to identify information that is frequently reported in each primary study such as sample size or age (Thompson, 1999). Thompson (1999) also noted that the characteristics of the instrument
will influence the variables that can be included in a RG study. For example, if an instrument has been translated into multiple languages, than the language of the instrument could be one of the instrument characteristics that is coded.

Dimitrov (2002) argued that one problem in RG studies is that some researchers may improperly code groups (e.g., gender). For example, Sawilowsky (2000a) critiqued Vacha-Haase’s (1998) RG study and noted that some of the independent variables were confounded; for example, gender was coded twice. Specifically, in Vacha Haase’s study gender was coded as all female or not, and it was also coded as having both males and females or only one gender (all males or females).

According to Dieckmann et al. (2009), unreliability in coding procedures of meta-analytic studies can add random variation to the analysis and reduce the reliability and power of the results. This problem can be addressed by using multiple trained raters and calculating interrater reliability or more specifically absolute interrater agreement for characteristics of the study and instrument that were coded. Dieckmann et al. recommend that authors conducting meta-analytic studies use multiple raters, describe a method of conducting interrater reliability, and report their interrater reliability estimate(s).

Analyzing and reporting results. After collecting, organizing, and coding data researchers must analyze their data and report their RG results.

Independence of reliability reports. Romano and Kromrey (2002) argued that an issue occurring in some RG analyses is that the reliability estimates analyzed do not represent independent observations. Lack of independence of observations includes analyzing estimates in the same statistical analysis from multiple subgroups, multiple
types of reliability, multiple subscales, and/or multiple instruments all derived from the sample study sample.

To address the concern regarding violations of independence, Romano and Kromrey (2009) conducted a Monte Carlo study that examined five different approaches for handling non-independence of reliability estimates: ignoring the problem, using the mean or median for each study, using one observation per study, or using a mixed-effects model. The results of their study showed that the type of approach used for handling non-independence of reliability estimates did not have a strong influence on the accuracy of the reliability results for the conditions that were simulated in their study. However, Romano and Kromrey (2009) noted that researchers should be careful when the intraclass correlation or dependency is large (their simulation included intraclass correlation values of .00, .01, .30, and .90). The results for all five treatments were similar when the bias in the mean estimates, root mean square error (RMSE) estimates, and confidence bandwidths were examined (although the results produced narrow bands). However, when the confidence band coverage was examined, the results suggested that calculating a mean of the reliabilities from each study gave better band coverage when compared to other methods (Romano & Kromrey, 2009). Additionally, Romano and Kromrey (2009) noted that none of the approaches they examined of handling a violation of independence were effective in producing accurate confidence intervals (CIs) in the bulk of conditions investigated.

Researchers sometimes include reliability estimates from multiple subgroups (e.g., boys and girls) in a single study. When RG researchers use subgroups as independent reliability estimates it can be a problem because of the dependency among
score reliabilities (Beretvas & Pastor, 2003). Both Rodriguez and Maeda (2006) and Beretvas and Pastor (2003) contend that the possible violation of independence must be addressed when multiple estimates of a reliability are reported from a single study.

Another issue of independence of reliability reports occurs when there are multiple types of reliability reported from one study. When conducting a RG study researchers need to decide on the type(s) of reliability estimates they want to examine in their RG study (e.g., internal consistency, test-retest). Sometimes RG researchers include multiple types of reliability in their studies. If authors do include multiple types of reliabilities in one RG study, then the literature recommends not using multiple types of reliability (e.g., internal consistency and test-retest) in a single analysis as different reliability estimates model different sources of measurement error (Beretvas & Pastor, 2003; Henson & Thompson, 2002; Rodriguez & Maeda, 2006) and violate basic meta-analytic principles (Dimitrov, 2002). Rather, different estimates of reliability should be examined in separate analyses, if there are sufficient numbers of each type of reliability being examined (Henson & Thompson, 2001).

A third issue regarding independence of reliability reports occurs when there are multiple subscales reported from the same instrument. Currently, researchers have not made recommendations regarding the issue of having multiple subscales for RG studies. However, as previously noted, both Rodriguez and Maeda (2006) and Beretvas and Pastor (2003) contend that the possible violation of independence must be addressed when multiple estimates of a reliability are reported from a single study (e.g., having multiple estimates from different subscales). Additionally, Cronbach (1951) noted: from the perspective of interpretability, the smallest part on which a score is attained should be
a set of items that have a substantial alpha and which are not able to be divided into smaller groups of items which themselves have a high alpha. Cronbach noted that these separate groups or subscales of items can sometimes be combined into an interpretable composite. Since subscales are individual units with their own reliability estimates, researchers should not combine multiple subscales in a single analysis.

Another issue regarding the independence of reliability reports occurs when researchers use different instruments of the same construct. Rodriguez and Maeda (2006) noted that a violation of the independence of reliability estimates transpires when several instruments are used to measure a certain construct and estimates from each instrument are derived from the same sample. Additionally, Borsboom and Mellenbergh (2002) argued that construct scores (which reflect the score on the attribute of interest) and “true” scores are different concepts. In practice, it is not possible to determine the construct score when using CTT methods; rather, researchers attain an observed score, which is the result of an investigative process. Since it is not possible to obtain construct scores in applied research, RG researchers should not combine multiple instruments in a single analysis (Rodriguez & Maeda, 2006).

**Weighting.** Dieckmann et al. (2009) noted that meta-analytic researchers must determine whether weighting by sample size is appropriate for their study. According to Romano and Kromrey (2002), one of the major controversies in RG studies concerns whether it is necessary for researchers to use weights in RG studies. They noted that the use of sample weights (weighting every reliability estimate by an estimate of its sampling error) is relatively rare in RG studies. Additionally, Romano and Kromrey (2002)
recommend that the use of sample weights is not necessary for simple descriptive applications.

Dieckmann et al. (2009) noted that most methodologists suggest weighting by precision (i.e., weighting each effect size by sample size or the inverse of the variance) when conducting meta-analyses. Additionally, Rodriguez and Maeda (2006) provided suggestions for conducting meta-analyses of coefficient alpha, and they suggested the use of weighting based on a function of the precision of each coefficient alpha value. Rodriguez and Maeda noted that usually the precision of the alpha value is established from its sampling distribution. Since each alpha value comes from a different study (e.g., different in characteristics such as sample size or group variability), each alpha value is estimated with a different level of precision.

Dieckmann et al. (2009) stated that there are some situations where weighting studies with larger sample sizes could produce misleading average effects due to confounding variables (e.g., studies with larger sample sizes used different methodologies than the other studies). Dieckmann et al. advised authors to determine whether sample size is confounded with other characteristics of the study sample before weighting by sample size.

**Homogeneity of variance.** Rodriguez and Maeda (2006) recommend that RG researchers conduct a test to examine the homogeneity of population of reliability estimates. When examining the homogeneity of population of reliability estimates, the null hypothesis is that the population estimate for each study is equal across all the studies. In other words, researchers examine whether the sample reliability estimates seem to be similar across all the studies.
Beretvas and Pastor (2003) indicated that a $Q$ statistic is frequently used to test the homogeneity of population correlations. However, one shortcoming of the $Q$ statistic is that it only shows the presence or the absence of heterogeneity (i.e., variability); it does not give the degree of heterogeneity (Huedo-Medina, Sánchez-Meca, Marín-Martínez, & Botella, 2006). Additionally, the results of a Monte Carlo simulation study by Harwell (1997) showed that a limitation of the $Q$ statistic is that it has poor power to detect heterogeneity among studies when small sample sizes are paired with larger ones in a meta-analysis. Harwell’s simulation study included five different sample size $N$s: 10, 20, 40, and 200 (which they considered a very large within-study sample).

Schmidt et al. (2009) found that the $Q$ test for homogeneity did not always correctly identify heterogeneity among correlations of effect sizes. However, a random-effects model did not miss the variability. Additionally, Beretvas and Pastor (2003) noted that researchers using mixed-effects models can examine whether there is a significant amount of variance between correlations, which is comparable to the $Q$ test.

**Power analysis.** The publication manual of the APA (2010) states that when researchers use inferential statistics, they must be aware of the statistical power considerations associated with testing hypotheses. For example, power considerations that relate to the probability of accurately rejecting the tested hypotheses, given a certain nominal alpha level, effect size, and sample size. Cafri, Kromrey, and Brannick (2009) recommend that researchers conduct an a priori power analysis prior to conducting meta-analytic studies, and this recommendation could easily be extended to RG studies as they fit under the meta-analytic framework. There are currently not any papers that discuss how to conduct power analyses for RG studies. However, Hedges and Pigott (2001)
provide a guide for meta-analytic researchers on how to compute statistical power for both fixed-effects and random-effects statistical analyses.

**Data transformation.** RG researchers disagree on whether it is necessary to transform reliability estimates prior to data analysis. Some authors (Henson & Thompson, 2002; Thompson & Vacha-Haase, 2000) have argued $r$-to-$z$ transformation is not necessary prior to conducting a RG analysis. Mason, Allam, and Brannick (2007) noted that if the reliability population values have a variance that is not zero, the transformation skews the distribution.

Multiple researchers (e.g., Henson & Thompson, 2002; Sawilowsky, 2000a) have argued that it is unnecessary to transform the internal consistency reliability estimates of coefficient alpha or $KR21$. On the other hand, Onwuegbuzie and Daniel (2000) contend that internal consistency reliability estimates should be transformed before conducting RG analysis. Onwueguzie and Daniel argued that “because internal consistency estimates are essentially a type of correlation coefficient, . . . the sampling distribution of the sample reliability estimate for all values of the theoretical reliability estimate other than zero is skewed” (p. 15) or asymmetric. Therefore, reliability estimates need to be transformed so that it has a sampling distribution that is approximately normal.

Beretvas and Pastor (2003) noted that many researchers have used Fisher’s (1928) $r$-to-$z$ transformation when performing meta-analyses of correlations to normalize the sampling distribution of $r$. Fisher defined the $r$-to-$z$ transformation as follows:

$$z = \left(\frac{1}{2}\right) \log_e \left(\frac{1 + r}{1 - r}\right)$$

where $z$ is the symbol for the transformed correlation; $\log_e$ is the natural logarithm, and $r$ is the correlation estimate.
Romano and Kromrey (2002) conducted a Monte Carlo study examining the potential influence of several factors on the validity of conclusions drawn from RG studies including \( r \)-to-\( z \) transformation. Their results show that for the majority of the conditions, the use of Fisher's \( r \)-to-\( z \) transformation led to a modest increase in the accuracy of the estimation of the population mean reliability. It is important to note that the Monte Carlo study design included different values for the true population reliability, and was not based upon a specific type of reliability estimate. Romano and Kromrey (2002) found transformations to be useful for RG studies when there were many primary studies, but small samples within each study, or when there was missing sample reliability estimates. However, Romano and Kromrey (2002) recommend that transformation is not necessary for simple descriptive applications.

**Data analysis.** Multiple researchers have provided recommendations concerning the data analysis of RG studies. RG researchers ought to make certain they are analyzing the data with the most appropriate models due to the amount of effort they put into collecting and coding data (Beretvas & Pastor, 2003). Warne (2008) stated that the first step in analyzing RG data for some researchers is to find the average reliability estimates. Feldt and Charter (2006) presented formulas for calculating the average reliability estimates.

Warne (2008) noted that the choice of data-analytic technique varies across researchers. Thompson and Vacha-Haase (2000) argued that RG studies do not have to use a single method of analysis and researchers can use a variety of analyses including Structural Equation Modeling (SEM), Analysis of Variance (ANOVA), or regression. Cousin and Henson (2000) stated that regression or other General Linear
Model (GLM) analyses should be used when conducting RG studies, but that when there is more than one subscale, it may be better to use multivariate analysis. Both Beretvas and Pastor (2003) and Rodriguez and Maeda (2006) contend that it is more appropriate to run multivariate mixed-effects or random-effects analyses or two mixed-effects univariate analyses when there are multiple types of reliabilities being analyzed and when researchers want to make inferences beyond the studies included in a RG study.

Wang (2002) suggested that Hierarchical Linear Modeling (HLM) is one way to conduct data analyses in RG studies. Wang stated that score reliability of an instrument relies upon many factors that make-up the instrument and individuals; therefore, factors should be introduced at multiple levels to describe these conditions and facilitate the generalization of the reliability. According to Raudenbush and Bryk (2002), HLMs are appropriate for analyzing meta-analytic data because data are hierarchically structured: participants are “nested” within studies; therefore, models should account for variation at both the participant and study level. Raudenbush and Bryk discussed two-level HLM models that researchers can use when conducting data analysis for meta-analytic studies. The level-1 model (or within-studies model) is also known as the unconditional model because no characteristics that predict effect sizes are in the equation. The level-2 model (or between-studies model) includes characteristics that predict the effect sizes. For additional details on using HLM for meta-analytic studies, interested readers can see chapter seven of Raudenbush and Bryk, or Wang (2002).

In their study of fixed- versus random-effects models in meta-analytic studies, Schmidt et al. (2009) applied random-effects procedures to 68 previously published meta-analytic studies which originally used fixed-effects analyses (e.g., regression). They used
two different random-effects models, one by Hedges and Vevea (1998) and another by Hunter and Schmidt (2004). The results of their study showed that the fixed-effects models typically underestimated the width of the CIs, which is problematic as this may lead to raised Type I error rates or false positives when researchers interpret CIs as significance tests. A Type I error is the error of rejecting a true hypothesis (Wiersma & Jurs, 2009). Additionally, Schmidt et al. noted that the circumstances under which fixed-effects analyses were appropriate for meta-analytic studies were limited (e.g., the primary studies included in the meta-analysis were nearly identical). Schmidt et al. noted that nearly identical studies are those that are “all literal or operational replications of each other” (p. 124).

**Characterizing variability.** RG researchers also need to decide on how best to present their results and characterize variability. Both Cousin and Henson (2000) and Warne (2008) recommend that fluctuations in reliability estimates should be examined descriptively, through a box and whisker plot or other graphical presentation. Box and whisker plots can also be used to compare scales if an instrument has multiple subscales (Warne, 2008).

Additionally, CIs can be used to characterize variability. Henson (2004) noted that combined CIs allow RG researchers to see the precision in their estimates. The publication manual of the APA (2010) stated that using CIs can be an effective method of reporting results and strongly recommended that researchers use CIs. APA contends that CIs are an excellent reporting method because they include information on both location and precision. Additionally, Onwuegbuzie and Daniel (2000) stated that RG researchers should report CIs around reliability estimates. Although the recommendation to use CIs
by both APA and Onwuegbuzie and Daniel were written for individual studies, these recommendations can be extended other studies including meta-analytic studies.

**Synthesis of Recommendations**

As you can see, authors have made suggestions for how RG researchers should collect studies, organize studies, code studies, analyze data, and report the results of RG studies. The purpose of this section is to synthesize these recommendations and divide them into three categories: essential, optimal, and controversial. These categories were created after the author read the recommendation articles and established an appropriate breakdown of the recommendations. When reading the articles, the author noticed a trend that some recommendations were supported by multiple research articles and other recommendations were supported by some authors and not supported by others. In the present study, essential RG recommendations are practices that all RG researchers should engage in no matter what year the study was conducted. Essential recommendations are those recommendations that many researchers support and which are currently considered best practice. Optimal recommendations are recommendations that are currently considered best practice, but as they are newer, it is not expected that older studies engaged in these practices. Controversial recommendations are recommendations that have some authors supporting the practice and others arguing against the practice.

There are multiple essential RG recommendations that pertain to the collection and organization of RG papers. For example, all researchers must conduct thorough searches for RG papers. In order to obtain the maximum number of relevant studies, RG researchers should use more than one search method (see e.g., Warne, 2008). Additionally, RG researchers should use more than one search term in order to find the
relevant studies (see e.g., Henson & Thompson, 2001). RG researchers must determine the criteria for inclusion of primary studies; they should not only include studies that were published in journal articles as this may contribute to the file-drawer problem (see e.g., Romano & Kromrey, 2002). Researchers should address the file-drawer problem by conducting a Fail-Safe N (see e.g., Howell & Shields, 2008), or using the KR21 formula if the item response scale system is dichotomous (see e.g., Cousin & Henson, 2000). Additionally, researchers can test for publication bias by using the funnel plot technique of testing for publication bias (see e.g., Light & Pillemer, 1984).

There are also essential RG recommendations pertaining to coding and reporting results. RG researchers should code both instrument characteristics and sample characteristics that may influence score reliability (see e.g., Thompson, 1999). RG researchers should also use multiple raters to code the RG studies and report at least one method of calculating interrater reliability (see e.g., Dieckmann et al., 2009). RG researches should use box and whisker plots to graphically present variability in their results (see e.g., Warne, 2008) or use CIs in their studies (see e.g., Henson, 2004).

There are some optimal RG recommendations pertaining to the analysis of RG studies. One issue that occurs in some RG studies is a lack of independence of reliability estimates. Henson and Thompson (2002) recommended that authors not combine multiple types of reliability. Additionally, Beretvas and Pastor (2003) recommended that authors not combine multiple subgroups in one analysis. Researchers who wrote articles before these publications may not have engaged in what is now considered best practice (e.g., researchers may combine multiple types of reliability in a single analysis). RG researchers should also conduct separate analyses when they have multiple subscales or
multiple instruments included in their study (see e.g., Rodriguez & Maeda, 2006). Additionally, RG researchers should examine the homogeneity of population correlations, this issue was not addressed by RG recommendation papers until 2003 by Beretvas and Pastor; therefore, researchers before this point in time may not engage in this aspect of what is now considered best practice. Finally, it is now considered best practice to conduct an a priori power analysis before conducting meta-analytic studies like RG. However, this is a relatively new recommendation in the RG literature; it was recommended by Cafri et al. in 2009, thus it was expected that the majority of researchers would not conduct an a priori power analysis.

Some RG practices are also controversial. For example, there is controversy regarding whether it is necessary to use weighting (Romano and Kromrey, 2002). However, researchers should use precision weights as long as sample size is not confounded with any other characteristics of the sample (Dieckmann et al., 2009). Another controversial aspect of RG practice concerns the transformation of reliability estimates. Multiple authors have debated whether it is necessary to transform reliability estimates, with some supporting the practice (see e.g., Onwueguzie & Daniel, 2000) and others opposing the practice (see see e.g., Henson & Thompson, 2002). However, it is optimal to not transform reliability estimates based upon the argument that if the reliability of population values have a variance that is not zero, the transformation skews the distribution (Mason et al., 2007) and other arguments previously discussed.

There is also controversy regarding the type of data analysis researchers should use when conducting RG studies. The current recommended practice would be for RG
researchers to use mixed-effects or random-effects analyses in their RG studies (see e.g., Beretvas & Pastor, 2003).

Limitations of RG Research

There are some limitations of RG research. As previously noted in the literature review section, one problem with RG studies is the limited amount of data that are available for inclusion because of the low reporting rates of reliability estimates in primary studies (Warne, 2008). Low reporting rates of reliability limits the generalizability of RG studies and constrains researchers’ understanding of how well an instrument measures a construct across different samples (Warne, 2008).

Warne (2008) noted that another limitation of RG research is that the most reported measure of reliability is Cronbach’s coefficient alpha. The predominance of the use of coefficient alpha indicates that researchers frequently only examine internal consistency reliability estimates (Warne, 2008).

Summary

First, this literature review examined reliability. This literature review also examined RG. Finally, RG recommendations were discussed and summarized. This chapter showed the importance of both reliability and RG. Additionally, areas where there were differences among the RG recommendations were also discussed. The next chapter will describe the methodology of the present study.
Chapter Three: Method

This chapter will explain the methods used in the present study the data that were used, the procedures, and data analyses. The aims of the present study include examining the current practices of researchers conducting RG studies and comparing how the current practices of RG researchers compare to RG recommendations.

Sample

Data consisted of RG papers identified by using the keywords reliability, generalization, meta-analysis, and combinations of these words in the PsycINFO and ERIC databases. The database search was limited to the years from 1998, when the seminal RG paper by Vacha-Haase was published, to 2010. Only studies that were published in English were included in this paper. Additionally, irrelevant studies (i.e., studies not related to RG) and studies that were duplicated in both databases (PsycINFO and ERIC) were not included in this paper. Initially, 490 studies were identified from searching the databases; there were 389 studies that were irrelevant, a repeat of a study from the other database or in a foreign language. The author read the recommendation papers and noted any applied RG studies that were cited by the recommendation authors, and attained any applied RG papers cited that were not found by the original searches.

An additional search was also made of Dissertation Abstracts using the same three keywords that were used to search for the initial RG papers. A comparison was made between the dissertations and theses found and the applied RG research papers to determine whether the authors later presented their results at a conference or published their dissertations or theses. The dissertations and theses were only coded by the author due to time constraints; the author did not want to require the other trained raters to code
these due to the amount of extra time it would take. Finally, the researcher listed as the contact on every RG paper was e-mailed to inquire whether they had any applied RG studies that were never published (which resulted in one additional RG study).

In this paper, both published and unpublished applied RG studies (i.e., studies that conducted a RG study) were included for coding, while studies providing recommendations or simulation results were incorporated into the RG recommendation section. When a study was identified as having been presented at a conference and later published, the most recent version of the study was used.

Two studies were removed from the data analyses of the present study. The studies by Vassar and Hale (2009) and Spector (2005) were removed from the analyses in the present study because the studies did not fit the previously mentioned criteria for a RG study as discussed by Vacha-Haase (1998). In her study, Spector only examined the reliability estimates reported in manuals. In their study, Vassar and Hale gave the reliability reporting practices in primary studies, examined trends in reliability reporting across time, and looked at whether the quality of the journal influenced reliability reporting practices.

Three applied RG studies (Bornmann, Mutz & Daniel, 2010; Helms, 1999; Mji & Alkhateeb, 2005) did not look at sources of variability in reliability estimates (which is one purpose of RG studies according to Vacha-Haase, 1998). These studies were not included in all of the analyses because some study characteristics that were coded (e.g., gender) were not relevant for these studies as their purpose was not to look at sources of variability and it would have been inappropriate to penalize authors for something that was not the purpose of a study. Specifically, Mji and Alkhateeb (2005) and Helms
(1999) were not included in the analyses of the following categories: coding instrument characteristics, coding sample characteristics, and interrater reliability for coding. Although the studies by Capraro and Capraro (2002) and Miller, Woodson, Howell and Shields (2009) did not conduct data analysis examining factors that influenced reliability, they both initially planned to examine predictors, but too few primary studies reported study characteristics.

The RG study by Bornmann et al. (2010) was removed from all the main analyses because it examined interrater reliability and it did not look at sources of variability in reliability estimates. This study also did not examine a particular instrument or group of instruments that examined the same construct like every other RG study. The study by Bornnammn et al. was individually examined because it was a contribution to the RG literature even though it was different then the other RG studies.

Overall, 64 applied RG studies were identified. See Appendix A for a complete list of applied RG studies. However, the majority of the analyses were conducted with 59 studies because the four dissertations/theses and the study by Bornmann et al. (2010) were excluded from the analyses (for the previously discussed reasons). Additionally, as previously noted, Mji and Alkhateeb (2005) and Helms (1999) were not included in all of the analyses.

Procedure

After identifying studies, the RG recommendation papers were first reviewed. The meta-analytic review article by Dieckmann et al. (2009) served as a foundation for recommendation characteristics to report in the recommendation section.
A coding scheme was created and was based upon the RG procedures suggested in the recommendation papers and the meta-analytic review article by Dieckmann et al. (2009). In addition to coding the content of the papers, the references of the papers were examined to see whether different translations of the instruments were used, and if papers were published in different languages. The references were also examined to determine if people included RG studies that were conference papers, books, and/or dissertations/theses.

**Interrater Reliability**

The raters created a coding sheet of the relevant variables for them to use when coding RG studies. See Appendix B for the coding sheet the raters used, and Appendix C contains the complete codebook. The four raters included a professor who specialized in measurement and three graduate students in an advanced measurement course in which the students studied reliability generalization. Before the raters began coding RG papers, the author trained the other raters by reviewing the coding sheets with them and clarified any questions they had concerning how to code the papers. The raters met multiple times to discuss how to code articles. In the present study, four raters coded the applied RG studies. The author and one other rater reviewed each of the applied studies, except for the dissertations and theses, which were only coded by the author. The four raters met multiple times to discuss the coding of the RG papers, any problems they encountered while coding, and to discuss suggestions for additional codes. Raters reviewed and re-coded papers after any changes were made to the coding (for example, after observing that multiple RG researchers contacted authors for additional RG studies, the variable was added to the code sheet).
In order to address interrater reliability, raters met to discuss and resolve any discrepancies in coding. If two raters disagreed and were not able to reach a resolution, a third rater was consulted until the coding disagreement was resolved. The interrater reliability was calculated by percent agreement. Initially, raters had an agreement rate of 97%, and after issues were resolved, raters reached 100% agreement.

**Data Analysis**

Once all applied RG papers were coded, they were compared to the recommended practices. As previously discussed in the literature review section (Chapter Two), RG recommendations were divided into essential, optimal, and controversial recommendations; this section will be organized by these three categories. This section will describe the analyses that were used to compare the RG practices to the RG recommendations.

In the results section, a variety of variables that were included by the RG studies in the data analyses is discussed. In the present study, the focus of the data analysis was on the most sophisticated data analysis that was completed by the authors of a RG study. For example, if a study conducted descriptive statistics regarding the variable gender, but did not include gender in the HLM analysis (which was conducted with other variables), then it was coded that the study did not include gender in the data analysis. Sometimes authors wanted to include certain variables in the data analysis, but were unable because of inconsistent reporting of the characteristics in the primary studies. Thus, some study authors reverted to less sophisticated data analysis methods.

The purpose of the present study was to survey the field descriptively. Additionally, it is important to note that the aim of this study was to examine the methods
used by researchers when conducting the RG studies; therefore, the results found by the RG studies were not examined. Furthermore, the purpose of this study was to examine trends across RG studies, not focus on individual studies that were exemplary or particularly poor.

**Data analysis of essential recommendations.** There are multiple RG recommendations that are included in the essential recommendation category. For example, RG researchers must first assemble the relevant primary studies when conducting a RG study. RG researchers can conduct a RG study with one instrument or with multiple instruments that examine the same construct, but each instrument is analyzed separately. Therefore, the number of instruments used in an RG study was tallied.

It has been recommended that RG researchers search electronic databases, the references of meta-analytic studies and contact well-known researchers who use the instrument of interest (see e.g., Thompson, 1999; Warne, 2008). Furthermore, researchers should use keywords (including abbreviations) that are broad enough to capture different forms of the instrument (Henson & Thompson, 2001). Descriptive statistics were used to compare the practices researchers have used for finding papers to the aforementioned recommendations. Specifically, the number of applied RG papers that found studies by examining review papers, searching electronic databases, or examining the references of meta-analytic studies were tallied. Additionally, the number of applied RG papers that found studies by using the title of the instrument(s), or abbreviations was tallied. Any additional methods of searching for papers or keywords used were also tallied.
Missing data is one problem facing RG researchers and researchers should address the issue by conducting a Fail-Safe $N$, or using the $KR21$ formula if the data are dichotomous (Howell & Shields, 2008; Kuder & Richardson, 1937; Orwin, 1983; Soeken & Sripusanapan, 2003). Additionally, researchers can also address the issue of missing data by testing for publication bias by using the funnel plot technique (Light & Pillemer, 1984). The number of RG applied papers that conducted a Fail-Safe $N$, used the $KR21$ formula, or used the funnel plot technique was tallied. Any additional methods studies used to address the problem of missing data or test for publication bias were also noted.

It is also essential for RG researchers to determine criteria for including primary studies. It was previously noted in the literature review section that meta-analyses usually only include published studies (Romano & Kromrey, 2002). In this study, whether researchers included conference papers or unpublished papers, book chapters, dissertations or theses, or journal articles was examined. Descriptive statistics were used to examine what types of studies RG researchers included in their analyses.

Coding is an essential part of RG studies. Researchers use two overarching categories when coding variables that influence score reliability: characteristics of the instrument and characteristics of the sample (Thompson, 1999). The instrument characteristic variables examined in this study included the response scale system of the items (i.e., whether the response scale system varied across studies, for example, some studies used a 5-point scale and other studies used a 4-point scale), who completed the instrument (e.g., self or someone else), and language of the instrument. The other instrument characteristics coded were the length of instrument (which was defined as coding the number of items on the instrument), and instrument forms. The variable
“included multiple forms” included instruments that had various forms where the wording was different. The variable “included multiple forms” also included instruments that were different lengths but the specific number of items was not coded as a variable (e.g., the studies coded long and short form, but did not code the specific number of items on the instrument). If a study included multiple instruments, but each instrument only included one form, then the study was identified as not including multiple forms.

The sample characteristic variables examined in this study included sample size, gender, racial or ethnic identity, age, and population type (e.g., patient or non-patient). Another variable that studies should code is the type(s) of reliability that the RG researchers used in their studies. Any additional sample characteristics that were coded were also noted. Descriptive statistics were used to summarize the variables that RG researchers included in their analyses.

Given the number of characteristics coded in each study, it is essential for RG researchers to have multiple raters to ensure paper characteristics are accurately coded (Dieckmann et al., 2009). Whether RG researchers conducted interrater reliability agreement was examined in this study. Additionally, if RG researchers did examine interrater reliability, the number of raters used and the type of analysis conducted (e.g., percent agree, Cohen’s Kappa or other type) was noted. Descriptive statistics were used to summarize interrater reliability practices in RG studies.

RG researchers must also decide how best to characterize the variability in the score reliabilities. Box and whisker plots can be used to show fluctuations in the reliability estimates (Warne, 2008). In this study, descriptive statistics were used to summarize researchers that used box and whisker plots. Additionally, another way to
examine variability in RG studies is with CIs. Henson (2004) recommended that RG researchers use CIs for their RG studies. Descriptive statistics were used to summarize RG practices regarding CIs.

**Data analysis of optimal recommendations.** There are multiple RG recommendations that are included in the optimal recommendation category such as the controversy concerning the independence of reliability reports. Romano and Kromrey (2002) noted that one issue that occurs in RG studies is that the reliability estimates do not represent independent observations. Lack of independence of observations includes analyzing estimates from multiple subgroups, multiple types of reliability, multiple subscales, and/or multiple instruments. Descriptive statistics were used to summarize whether violations lack of independence of observations was observed in RG studies.

When examining independence issues, the methods that researchers used to handle a lack of independence were also examined. Additionally, the methods that more recent applied RG papers used to handle a lack of independence regarding subgroups and types of analyses were compared to older papers to determine whether there were any changes in RG practices over time.

In their analysis, RG researchers can use tests to examine the homogeneity of population reliability estimates (Beretvas & Pastor, 2003). If RG researchers did test for the homogeneity of population coefficients the type of test used (e.g., $Q$ statistic) was recorded in this study. Any other methods of conducting tests for the homogeneity of population correlations were also noted. Descriptive statistics were used to summarize researchers’ use of tests of homogeneity of population coefficients. Additionally, A
figure will be used to track trends in conducting tests of homogeneity of variance over time and to examine if there was a change after the paper by Beretvas and Pastor in 2003.

Cafri et al. (2009) suggested that researchers conduct an a priori power analysis before conducting meta-analytic studies. In this study, descriptive statistics were used to examine whether RG researchers conducted an a priori power analysis.

**Data analysis of controversial recommendations.** There are multiple RG recommendations that are debated amongst RG researchers and are included in the controversial category. For example, one controversial recommendation is that RG researchers should use weights during their analyses. Descriptive statistics were used to examine whether RG researchers were using weights and if so, what type of weights were used (e.g., sample size). Any other methods of weighting that RG researchers used were also noted. Romano and Kromrey (2002) noted that the use of sample weighting was rare in RG studies. A figure will be used to track trends in weighting over time and to examine if there was a change after the paper by Romano and Kromrey in 2002.

Another issue where there is a debate amongst RG researchers is whether it is necessary to transform reliability estimates. In this study, descriptive statistics were used to summarize whether or not researchers transformed the data prior to conducting analyses. Since the $r$-to-$z$ transformation debate spans across years, a clear cut date was not used to compare RG studies across time.

Multiple researchers have given recommendations concerning how RG researchers should conduct their data analyses. The most sophisticated type of data analysis conducted was coded in this study (i.e., if researchers conducted descriptive statistics and random-effects analyses, then random-effects was coded as the type of data
Descriptive statistics were used to summarize the type of data analysis used by RG researchers. It is important to note that some researchers stated that they were limited in the type of data analysis they could conduct due to low reports of reliability in the primary studies. Thus, the number of studies who reported that their data analysis was limited due to reliability reporting was also summarized.

**Data analysis of other RG practices.** There were some RG practices that researchers did not make specific recommendations regarding; these RG practices are included in the present section. The number of studies that gave a reason for selecting their instrument was tallied and descriptive statistics were used to summarize the reasons studies given for selecting an instrument. Additionally, the number of primary studies that reported reliability and the total number of reliability estimates used in each study was tallied. Finally, the number of reliability estimates reported in the primary studies was also tallied.

**Summary**

This chapter examined the methods used for the present study. The data, procedure, and analyses of the present study were discussed. The results of the present study are presented in the following chapter.
Chapter Four: Results

The purpose of the present study was to examine the current practices of researchers conducting RG studies and to compare these practices to a variety of RG recommendations, which were previously discussed in the literature review. As previously mentioned in the introduction section, the purpose of the present study was to answer the following research questions: (a) What are the current practices of researchers conducting RG studies? (b) How do the current practices of RG researchers compare to the RG recommendations? The results section includes data on adherence to essential, optimal, and controversial recommendations; it also includes an “other” category, which includes study variables that were relevant, but did not fall into one of the preceding categories. Additionally, it includes a brief review of the results of other RG studies that were not included in the main analysis. In the present study, CIs are reported for the results; the formula used was a sample confidence interval for proportions (see e.g., Glass & Hopkins, 1996).

Essential Recommendations

There was a variety of essential RG recommendations. Figure 5.1 summarizes the percentage of studies that followed the essential RG recommendations.

Instruments coded. The number of instruments that studies coded varied from one to 24. The majority of studies (81%) coded one instrument, 95% CIs [.35, .61].

Searching the literature. RG studies used different keywords when searching for papers including the title of the instrument (73%), acronyms or abbreviations (53%), the construct of interest (32%), and/or the scale author (20%), 95% CIs [.62, .84], [.40, .66], [.20, .44], and [.10, .30], respectively. Additionally, 12% of studies used keywords other
than the ones previously mentioned (e.g., reliability), 95% CIs [.04, .20]. The total of the aforementioned percentages adds up to more than 100% because some studies used more than one keyword when searching for studies. The majority of studies (66%) used two or more keywords when searching for papers, 95% CIs [.54, .78].

Most studies searched for papers using PsycINFO or PsychLit (86%), 95% CIs [.77, .95]. Additionally, 32% of studies searched for papers using ERIC, 95% CIs [.20, .44]. Some RG studies also examined review papers (9%) and the references of RG studies (25%), 95% CIs [.02, .16] and [.14, .36], respectively. Additionally, nine percent of studies contacted authors for RG studies, 95% CI [.02, .16]. The majority of studies (56%) also used other methods including search engines such as Pubmed and Academic Search Elite, 95% CIs [.43, .69]. The total percentage for searching for papers equals more than 100% because many studies used multiple methods to search for papers. The majority of studies (71%) used two or more methods to search for studies used in a RG study, 95% CIs [.59, .83].

Criteria for including studies. The majority of RG studies (95%) included journal articles/published papers in their study, 95% CI [.89, 1.01]. RG studies also included conference papers/unpublished papers (29%), books/book chapters (25%), and dissertations/thesis (31%), 95% CIs [.17, .41], [.14, .36], and [.19, .43], respectively. Additionally, 9% of studies included papers written in languages other than English, 95% CI [.02, .16]. The total for inclusion criteria sums to more than 100% because some studies included multiple entities in their analyses. Furthermore, 49% of RG studies included only journal articles, 95% CIs [.36, .62].
**Missing data.** The bulk of studies (88%) did not check for publication bias or file-drawer bias, 95% CIs [.80, .96]. Of the seven studies that checked for publication bias or the file-drawer problem, one study (14%) used Begg’s test (Begg & Mazumdar, 1994), and one study (14%) used only Fail-Safe $N$ for RG studies (Howell & Shields, 2008), 95% CIs [-.12, .40] and [-.12, .40], respectively. Additionally, one study (14%) used Fail-Safe $N$ by Soeken and Sripusanapan (2003), one study (14%) used the Fail-Safe $N$ by Soeken and Sripusanapan and a funnel plot, and three studies (21%) used Orwin's (1983) Fail-Safe $N$ for effect size, 95% CIs [-.12, .40], [-.12, .40], and [-.09, .51], respectively.

One way to address the issue of missing data is for studies to include additional reliability estimates in their study. Eight RG studies (14%) indicated that they had e-mailed the authors of studies that used their instrument of interest to request the reliability information from the original study, 95% CIs [.05, .23]. Additionally, some RG studies used the $KR21$ formula (Kuder & Richardson, 1937) to generate reliability information when the instrument of interest used a dichotomous scale. In the present study, 23 RG studies had at least one scale that was dichotomous. Of these, three studies (13%) used $KR21$ to generate reliability information, 95% CI [-.07, .27]. The present study did not code whether the response scale system was dichotomous or polytomous. Therefore, the percentage of studies that used the $KR21$ formula included all studies and not only those that used a dichotomous scale.

**Types of reliability.** RG studies used multiple types of reliability estimates. The most frequently used type of reliability estimate was coefficient alpha (98%), 95% CIs, [.94, 1.02]. Additionally, 3% of studies used interrater reliability, 24% of studies used
KR20/KR21, 10% of studies used split-half reliability, 23% of studies used test-retest reliability, and there were not any studies that used parallel forms reliability, 95% CIs [-.01, .07], [.13, .35], [.02, .18], [.12, .34], and [.00, .00], respectively. Additionally, some studies discussed that they dropped a type of reliability estimate from the data analysis for various reasons. Such as there being a limited number of reliability estimates reported in the primary studies (seven [12%] of 59 studies) or because they wanted to avoid problems associated with generalizing across different estimates of reliability (2 [3%] of 59 studies), 95% CIs [.04, 20] and [-.01, .07], respectively.

**Instrument characteristics that were coded.** The results showed that the majority of studies (90%) coded at least one characteristic of the instrument, 95% CIs [.82, .98].

The response scale system of the instrument(s) did not vary for a preponderance of instruments used in RG studies (67%), 95% CIs [.55, .79]. Of the 19 studies that had response scale systems that varied, five of the studies (26%) did not code for response scale system, 95% CIs, [.06, .46]. The studies that did code whether the response scale system varied had multiple ways that they coded this variable (e.g., Likert, other type of response scale system, or whether a four- or five-point response scale system was used). Of the 14 studies that coded type of response scale system, 71% of studies included type of response scale system in their data analysis, 95% CIs [.47, .95].

Who completed the instrument, (e.g., the individual completed the instrument or someone else completed it) did not vary in the majority of the studies (86%), 95% CIs [.77, .95]. Of the eight studies that varied in who completed the instrument, two of these studies (25%) did not code this variable in their study, 95% CIs [-.05, .55]. The studies
that coded whether who completed the instrument had multiple ways that they coded the variable such as whether it was completed by “self” or “other.” Of the six studies that coded who completed the instrument, 83% of the studies included the variable in the data analysis, 95% CIs [.53, 1.13].

The bulk of studies (54%) did not use multiple translations of the instrument(s), (e.g., including both English and Spanish versions of the instrument), 95% CIs [.41, .67]. Of the 26 studies that included multiple translations of the instrument in the RG study, two (4%) of the studies did not code this variable in the study, 95% CIs [-.04, .12]. Of the 24 studies that coded whether there were multiple translations of the instrument(s), 96% of the studies included the language of the instrument in the data analysis, 95% CIs [.88, 1.04].

The majority of studies (72%) included multiple forms of at least one instrument that was included in the RG study, 95% CIs [.60, .84]. Of the 41 studies that included multiple forms of any instrument, 13 of the studies (32%) did not code whether there were different forms in their study, 95% CIs [.18, .46]. Of the 28 studies that coded multiple forms of any instrument, 82% of the studies included the form of the instrument in the data analysis, 95% CIs [.68, .96].

The bulk of studies (77%) included instruments of varying lengths (this variable included both studies that had multiple instruments of different lengths and studies that had one instrument with shorter and longer versions), 95% CIs [.66, .88]. Of the 44 studies that had instruments of varying lengths, 25 (57%) did not code the number of items, 95% CIs [.42, .72]. Of the 28 studies that coded whether instruments varied in length, 84% included the number of items in the data analysis, 95% CIs [.70, .98].
A preponderance of studies (52%) did not code for standard deviation or variance, 95% CIs [.39, .65]. Of the 27 studies that coded standard deviation or variance, 85% included it as a predictor in the data analysis, 95% CIs [.72, .98].

**Sample characteristics that were coded.** The results showed that almost all RG researchers (96%) coded at least one sample characteristic, 95% CIs [.90, 1.01]. The majority of studies (84%) coded sample size as one of the variables in their study, 95% CIs [.74, .94]. Of the 48 studies that coded sample size, 66% included sample size as a predictor in the data analysis, 95% CIs [.53, .79]. Additionally, some studies coded sample size and used it for weighting, but did not include sample size as a predictor.

The bulk of studies (84%) coded gender as one of the variables in their study. The studies that coded gender coded it in a variety of ways, 95% CIs [.74, .94]. For example, some studies coded gender by groups (e.g., males, females, mixed group), but other studies coded gender as percent male, percent female, or percent of the majority. Of the 48 studies that coded gender, the majority (88%) included it in the data analysis, 95% CIs [.79, .97].

Many studies (49%) coded racial or ethnic identity as one of the variables in their study, 95% CIs [.36, .62]. The studies that coded racial or ethnic identity coded it in a variety of ways. For example, some studies grouped the variable (e.g., Caucasian or African Americans), other studies coded racial or ethnic identity as percent Caucasian or percent of the majority. Of the 28 studies that coded racial or ethnic identity, the bulk of studies (82%) included it in the data analysis, 95% CIs [.68, .96].
The majority of studies (81%) coded the participants’ age as a variable in their study, 95% CI [.71, .91]. Of the 46 studies that coded age, the majority (59%) used the mean age of the participants, 95% CI [.45, .73]. Some studies also coded age by grouping participants (30%), 95% CI [.17, .43]; the groupings differed across the RG studies (e.g., adolescents or adults). Of the 46 studies that coded age, the majority of studies (87%) included it in the data analysis, 95% CI [.77, .97].

Most studies (61%) coded population type as one of the variables, 95% CI [.48, .74]. Population type was usually broken down into groups such as clinical or non-clinical, or incarcerated or not incarcerated. Some RG studies also defined population type by student status (e.g., students or non-students). Given that population type was a convoluted variable, no additional analyses were conducted for the present study.

**Interrater reliability.** The majority of RG studies (70%) did not calculate interrater reliability for the coding in their studies, 95% CI [.58, .82]. Of the 17 studies that calculated interrater reliability, they all used two or three raters, and the majority (71%) used percent agreement to calculate interrater reliability, 95% CI [.49, .93]. Additionally, 18% of studies used Cohen’s kappa, 6% of studies used intra-class correlation, and 24% of studies did not report the method of calculating interrater reliability, 95% CI [-.003, .26], [-.05, .17], and [.04, .44], respectively. The above percentages for methods of calculating interrater reliability add up to more than 100% because three studies used more than one method of calculating interrater reliability.

**Characterizing variability.** The results showed that the majority of studies (64%) did not use a box and whisker plot to summarize their results, 95% CIs [.52, .76]. Additionally, the majority of studies (63%) did not report CIs, 95% CI [.51, .75].
Summary of essential recommendations. The results demonstrated that RG researchers used a variety of keywords, most researchers used more than one keyword when searching for RG papers, and studies tended to use multiple ways to search for papers. Although researchers included a variety of studies including journal articles and conference papers, many researchers only included journal articles in their studies. Additionally, most RG researchers did not follow the RG recommendation to check for publication bias. Furthermore, RG researchers used a variety of types of reliability estimates, and the vast majority included coefficient alpha in their data analysis. RG researchers coded a variety of instrument characteristics including the response scale system, who completed the instrument, language of instrument, instrument form, instrument length, and standard deviation or variance. RG researchers also coded a variety of sample characteristics including sample size, gender, racial or ethnic identity, age, and population type. However, most RG researchers did not follow the RG recommendation of calculating interrater reliability for the coding of their studies. Finally, the majority of studies did not follow the recommendation of using a box and whisker plot, and the bulk of studies did not follow the recommendation of reporting CIs.

Optimal Recommendations

There was a variety of optimal RG recommendations. For a brief overview of the percentage of studies that followed the optimal RG recommendations, see Figure 5.2

Independence issues. Lack of independence of observations includes analyzing estimates from multiple subgroups, multiple types of reliability, multiple subscales, and/or multiple instruments.
**Multiple subscales.** Most studies (59%) reported reliability for more than one subscale, 95% CIs [.46, .72]. Of the 35 studies that reported reliability for more than one subscale, the majority of studies (89%) conducted separate analyses for the different subscales, 95% CIs [.79, .99]. Additionally, some studies also combined multiple subscales in one analysis, dummy coded subscales, or chose to use only one subscale.

**Multiple instruments.** The majority of studies (80%) did not use multiple instruments in their RG studies, 95% CIs [.70, .90]. Of the 12 studies that used multiple instruments, six conducted separate analyses of the instruments, 95% CIs [-.07, .19]. Additionally, some studies also combined multiple instruments in one analysis or dummy coded different instruments.

**Multiple types of reliability.** The bulk of studies (64%) did not include multiple types of reliability in their analyses, 95% CIs [.52, .76]. Of the studies that included multiple types of reliability in their analyses, (62%) conducted separate analyses for different types of reliability, 95% CIs [.41, .83]. There were not any trends over time regarding combining multiple types of reliability in one analysis; four out of the 20 studies that included multiple types of reliability combined them in analysis, and each case occurred in a different year. Some studies also used only one type of reliability, or dummy coded type of reliability during analyses.

**Multiple subgroups.** One issue of independence occurred when there were multiple subgroups (e.g., males and females). Overall, thirty-nine studies (66%) used both subgroups and whole groups in their analyses, 95% CIs [.54, .78]. Of the RG studies that used subgroups, the majority of studies (95%) included multiple subgroups in one analysis, 95% CIs [.86, 1.02]. There were not any trends over time, as the vast
majority of papers combined multiple subgroups in one analysis. Additionally, the bulk of studies (81%) did not justify using multiple subgroups in one analysis, 95% CIs [.68, .94]. There were seven studies (19%) that justified using multiple subgroups in one analysis, 95% CIs [.06, .32]; all of these studies contended that all of the subgroups that were included were independent samples. Additionally, there were two studies (5%) that used HLM to handle the issue of having multiple subgroups in one analyses, 95% CIs [-.02, .12].

**Power analysis.** The majority of studies (95%) did not conduct an a priori or post-hoc power analysis, 95% CIs [.89, 1.01].

**Homogeneity of variance.** Overall, most studies (80%) did not conduct tests for homogeneity of variance 95% CIs [.70, .90]. Figure 5.3 gives the percentage of studies over time that conducted a test for homogeneity of variance. The results show an increase in the number of studies conducting tests of homogeneity of variance over time. Additionally, it is noteworthy that there was only one RG study in 1999, which made achieving 100% of studies conducting the test for homogeneity of variance easier than in years with more RG studies.

Of the studies that conducted tests for homogeneity of variance, most studies used the $Q$ statistic (67%), 95% CIs [.40, .94]. Additionally, there was one study (8%) from 2004-2010 that used a HLM random-effects test to test for homogeneity of variance, 95% CI [-.07, .23]. Additionally, four studies (33%) used other tests of homogeneity of variance, 95% CI [.06, .60]. The total number of tests of homogeneity of variance used sums to 13 instead of 12 because there was one study that used two different tests of homogeneity of variance.
Summary of optimal recommendations. The results showed that the majority of studies followed the recommendation of conducting separate analyses when there were issues of independence regarding subscales, instruments, or reliability types. However, the majority of studies included multiple subgroups in the same analysis. The vast majority of studies did not follow the recommendation of conducting an a priori power analysis. Finally, most studies did not follow the recommendation of conducting tests for homogeneity of variance.

Controversial Recommendations

The three controversial recommendations included whether researchers should transform the data, whether they should use a weighting approach, and what type of data analysis should be used in RG studies. For a brief overview of the percentage of studies that engaged in controversial RG recommendations, see Figure 5.4.

Data transformation. Most studies (62%) did not transform data before conducting analyses, 95% CIs, [.50, .74].

Weighting. Overall, most studies (75%) did not use a weighting approach, 95% CIs, [.64, .86]. Of the 15 studies that used a weighting approach, sample size was the most frequently used weighting approach for the studies (53%), 95% CIs [.28, .78]. Additionally, 33% of studies used the inverse variance weight, and two studies (13%) used other weighting methods, 95% CIs [.09, .57], and [.04, .30] respectively. Figure 5.5 gives the trend in use of a weighting approach over time. The figure shows that in general the number of studies that the number of studies increases after 2002 (which was the date of the recommendation paper by Romano and Kromrey).
Data analysis. The majority of studies (88%) conducted a fixed-effects analysis, 95% CIs [.80, .96]. Additionally, 3% of studies used descriptive statistics, 18% used correlations, 11% used random-effects analysis, and 2% of studies used other data analytic techniques, 95% CIs [-.01, .07], [.08, .28], [.03, .19], and [-.02, .06], respectively. Although descriptive statistics (e.g., Means, frequencies) and correlations fall under the category of fixed-effects, the decision was made to separate out these two types of analyses for illustrative purposes. Additionally, some studies (17%) used a different data analytic technique than they planned to employ, but stated that they were limited by the small number of reliability estimates reported in the primary studies, 95% CIs, [.07, .27].

Summary of controversial recommendations. The results showed that most studies did not transform the data before conducting analyses. Additionally, the majority of studies also did not use a weighting approach. Finally, the majority of studies conducted fixed-effects analyses.

Other RG practices

This section consists of multiple variables including how researchers selected an instrument, and the number of papers and reliability estimates that were given in RG studies.

Choosing an instrument. Most studies (86%) provided a rationale for selecting the instrument they used in the study, 95% CIs [.77, .95]. Of the studies that provided a rationale for selecting the instrument they used, the most frequent response given by 82% of studies was that the instrument was the most popular instrument used to measure the construct, or a widely used instrument, 95% CIs, [.72, .92]. Additionally, 12% of studies
stated that they wanted to find which instrument had the best reliability among several instruments, and 7% of studies noted that the selected the instrument because it was a superior instrument of a construct, 95% CIs [.04, .29] and [.005, .14], respectively. The total percentage adds up to greater than 100% because one study gave multiple instrument rationales.

**Number of papers and reliability estimates.** The number of primary studies (i.e., studies conducted by researchers who use the instrument[s] of interest) that reported reliability estimates varied from 5 to 215 and the median was 41 ($M = 86.24$, $SD = 177.42$, 25th = 23, 50th = 41, 75th = 90). Figure 5.6 provides a stem and leaf plot of the number of primary studies that reported reliability estimates. The total number of reliability estimates used in the RG data analyses ranged from 10 to 2,207 and the median was 113 ($M = 193.80$, $SD = 309.15$, 25th = 48, 50th = 113, 75th = 215. Figure 5.7 provides a stem and leaf plot of the number of reliability estimates used in the RG data analyses.

The majority of RG studies (80%) gave details about how reliability was reported in the primary studies, 95% CIs [.70, .90]. Overall, only 24% of primary studies reported reliability for their sample, 95% CIs [.13, .35]. The majority of primary studies (55%) did not mention reliability or report reliability estimates, 95% CIs, [.42, .68]. Additionally, 5% of primary studies mentioned reliability, but did not cite values, 9% of primary studies did reliability induction, and 7% of studies used other methods of reporting reliability in primary studies, 95% CIs, [-.006, .11], [.02, .16], and [.005, .14], respectively.
Summary of other aspects of the studies that were coded. The results demonstrated that the majority of studies gave a rationale for why they selected the instrument they used. There was a wide range of both the number of primary studies and the number of reliability estimates that were included in RG studies. Finally, the results also showed that the majority of primary studies did not mention reliability or report reliability estimates.

Other RG Studies

As previously mentioned in the method section, five applied RG studies were not included in the main analyses (four dissertations and theses and the study by Bornmann et al., 2010). The results of Bornmann et al. (2010) were similar to those of other studies. They followed most of the essential recommendations. The researchers used more than two keywords and more than two methods when searching for papers, included journal articles and other types of studies, checked for publication bias, included interrater reliability estimates, and reported CIs. However, they did not do interrater reliability for coding or use a box and whisker plot. Some characteristics coded in other RG studies were not relevant for this study (e.g., instrument characteristics), but the researchers did code relevant characteristics for their study such as the interrater reliability method used.

Bornmann et al. (2010) followed the optimal recommendation to conduct a test for homogeneity of variances. However, the researchers did not conduct an a priori power analysis. In regards to the controversial recommendations, Bornmann et al. transformed the data and conducted random-effects analyses; however, they did not use a weighting approach.
The results of the dissertations and theses were also similar to other applied RG studies. The studies typically followed the essential recommendations. All four studies used at least two methods of searching for studies, most studies used at least two keywords when searching for RG studies, only one study checked for publication bias, and the studies used a variety of types of reliability estimates. Additionally, all four studies coded both instrument and sample characteristics, one study used a box and whisker plot, and half of the studies reported CIs. However, all four studies only included journal articles, and none of the studies calculated interrater reliability.

Some of the optimal recommendations were followed in the dissertations and theses. Whenever studies included multiple subscales, instruments, or types of reliability, the researchers conducted separate analyses of the aforementioned variables; however, the studies that used subgroups combined the subgroups in one analysis. Additionally, none of the dissertations and theses conducted an a priori power analyses, and one study conducted a test for homogeneity of variance. In regards to the controversial RG recommendations, one study transformed the data, half of the studies used a weighting approach, and all four studies used fixed-effects analyses. Additionally, all four studies gave a rationale for selecting instrument(s).

Summary of the Results

Overall, the results showed that RG researchers engaged in a variety of practices. The results demonstrated that at times researchers followed recommendations, and other times they did not follow RG recommendations.
**Essential recommendations.** In general, RG researchers followed the essential recommendations. However, most studies did not check for publication bias, calculate interrater reliability for the coding of papers, or report CIs.

**Optimal recommendations.** Overall, RG researchers did not follow the optimal recommendations. The majority of RG researchers followed the recommendation of conducting separate analyses when there were issues of independence regarding subscales, instruments, or reliability types; however, the vast majority of studies included multiple subgroups in the same analysis. Furthermore, the majority of studies did not follow the recommendations of conducting an a priori power analysis, or conducting tests for homogeneity of variance.

**Controversial recommendations.** In regard to controversial recommendations, results showed that most studies did not transform data before conducting analyses or use a weighting approach. Additionally, the majority of studies conducted fixed-effects analyses.

**Other RG practices.** The results demonstrated that RG researchers typically gave a rationale for why they selected the instrument they used. Researchers also gave the number of primary studies included in their studies, and the number of reliability estimates used in their studies.

**Other RG studies.** The results showed that the results found in the RG study conducted by Bornmann et al. (2010) and the four dissertations and theses were similar to other applied RG studies.
**Figure 5.1.** Percentage of studies that followed essential recommendations. Search Term = used two or more keywords; Search Method = used two or more search methods; journal articles = did not only include journal articles; Publication Bias = checked for publication bias; Instrument Characteristic = Coded at least one instrument characteristic; Sample Characteristic = Coded at least one sample characteristic; Interrater Reliability = calculated interrater reliability; Box and Whisker = used a box and whisker plot; Confidence Intervals = used confidence intervals
Figure 5.2. Percentage of studies that followed optimal recommendations. In regards to issues of independence, the percentages only include the relevant studies that had multiple subscales, instruments, types of reliability, or subgroups. Subscale = conducted separate subscale analyses; instrument = conducted separate instrument analyses; reliability type = conducted separate reliability type analyses; Subgroup = conducted separate subgroup analyses; Power Analysis = conducted a power analysis; Homogeneity = conducted a test for homogeneity of variance.
**Figure 5.3.** Percentage of studies over time that conducted a test for homogeneity of variance.

**Figure 5.4.** Percentage of studies that engaged in controversial practices.
Figure 5.5. Percentage of studies over time that used a weighting approach.

<table>
<thead>
<tr>
<th>Steam</th>
<th>Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>12234477</td>
</tr>
<tr>
<td>2</td>
<td>0112333556</td>
</tr>
<tr>
<td>3</td>
<td>000034778</td>
</tr>
<tr>
<td>4</td>
<td>1134579</td>
</tr>
<tr>
<td>5</td>
<td>236</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>58</td>
</tr>
<tr>
<td>8</td>
<td>09</td>
</tr>
<tr>
<td>9</td>
<td>01345</td>
</tr>
<tr>
<td>10</td>
<td>26</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>7</td>
</tr>
<tr>
<td>14</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>3</td>
</tr>
</tbody>
</table>

Figure 5.6. Number of primary studies that reported reliability estimates. There was one extreme study that reported 215 reliability estimates.
<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>024</td>
</tr>
<tr>
<td>2</td>
<td>22246</td>
</tr>
<tr>
<td>3</td>
<td>4556</td>
</tr>
<tr>
<td>4</td>
<td>088</td>
</tr>
<tr>
<td>5</td>
<td>189</td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>37</td>
</tr>
<tr>
<td>8</td>
<td>99</td>
</tr>
<tr>
<td>9</td>
<td>0455</td>
</tr>
<tr>
<td>10</td>
<td>335</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>68</td>
</tr>
<tr>
<td>13</td>
<td>8</td>
</tr>
<tr>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
</tr>
<tr>
<td>16</td>
<td>15</td>
</tr>
<tr>
<td>17</td>
<td>468</td>
</tr>
<tr>
<td>18</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>12</td>
</tr>
<tr>
<td>20</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>556</td>
</tr>
<tr>
<td>22</td>
<td>9</td>
</tr>
<tr>
<td>23</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>2</td>
</tr>
<tr>
<td>26</td>
<td>7</td>
</tr>
<tr>
<td>27</td>
<td>2</td>
</tr>
</tbody>
</table>

*Figure 5.7. Number of reliability estimates used in the RG data analyses. Additionally, there were nine studies that reported that they used over 300 reliability estimates; these studies reported 315, 322, 328, 388, 450, 489, 680, 810, and 2,207 reliability estimates in the RG data analysis.*
Chapter Five: Discussion

The present study examined the recommendations that have been made for conducting RG studies, the current practices of researchers conducting RG studies, and examined how the current practices of RG researchers compared to the RG recommendations. The overarching hypothesis was that there would be differences between the RG recommendations and the current practices of applied RG researchers. Applied RG researchers followed some of the recommendations (e.g., RG researchers examined both sample characteristics and instrument characteristics that influenced reliability estimates). Yet there were some recommendations that RG researchers did not follow (e.g., the majority of researchers did not conduct an a priori power analysis).

Meta-analytic studies have a longer history than RG studies; however, as previously discussed in the introduction section, RG studies are a type of meta-analyses and RG researchers can look to the meta-analytic literature for reviews of meta-analytic practices (see e.g., Ahn et al., 2012; Dieckmann et al., 2009, & Geyskens et al., 2009). Additionally, Vacha-Haase and Thompson (2011) conducted a review of typical RG practices. However, this study was the first that examined whether RG researchers were following recommendations that have been made about conducting RG studies.

Although the years for both the present study and the study by Vacha-Haase and Thompson (2011) spanned from 1998 to 2010, the present study included 64 RG studies, whereas Vacha-Haase and Thompson’s article included 47 studies. There are several possible reasons for this difference. For example, the present study and the study by Vacha-Haase and Thompson had different methods of searching for RG studies. In addition to the methods used by Vacha-Haase and Thompson, the present study contacted
the authors of RG studies to gain additional RG studies. Additionally, when searching for RG articles the present study included the keyword meta-analysis in addition to the terms that were used by Vacha-Haase and Thompson. Furthermore, the present study also included a search of the Dissertation Abstracts database in addition to the databases used by Vacha-Haase and Thompson when searching for studies. After examining the references of the study by Vacha-Haase and Thompson it was determined that only journal articles were listed as RG studies included in their review whereas the present study included a variety of types of RG studies.

Although some of the RG practices examined were the same in the present study and in the study by Vacha-Haase and Thompson (2011), there were some differences. For example, the present study examined ways RG researchers addressed the issue of missing data, methods and keywords authors used when searching for studies, and whether studies used multiple coders and computed interrater reliability for coding.

Examining differences in RG practices by researchers is important because procedural choices such as how to retrieve, code, and analyze information can affect the conclusions drawn from meta-analytic studies (Schmidt et al., 2009) like RG. The following section includes a discussion of the recommendations that were followed by RG researchers, the recommendations that were not followed, and controversial recommendations. A discussion of the problem of the lack of reporting of reliability estimates in primary studies, other RG studies not included in the primary analyses, and an exemplary RG study are also included. Finally, recommendations for conducting RG studies, the importance of the present study, and future directions are discussed.
Recommendations That Were Followed

The results showed that many of the RG recommendations were followed by RG researchers. The following section includes a discussion of both essential and optimal recommendations that were followed by RG researchers.

Essential recommendations. The results showed that authors frequently followed the essential recommendations. For example, Henson and Thompson (2001) recommended that researchers use broad keywords, and the results showed that authors typically used at least two keywords when searching for papers. Additionally, Henson and Thompson (2002) stated that databases such as PsycINFO can help researchers find studies. The results showed that many authors used two or more methods of searching for RG studies. These findings are consistent with the meta-analytic study by Dieckmann et al. (2009) who reported that the majority of researchers (88%) used computer searches to find articles; additionally 27% of authors used three different methods in their study.

According to Romano and Kromrey (2002), meta-analyses are typically performed using only published studies. Although the majority of authors included a variety of studies, such as journal articles and conference papers, many RG researchers only included journal articles in their studies. However, Dieckmann et al. (2009) found that 53% of meta-analyses included unpublished studies.

One of the purposes of RG studies is to examine the sources of variability in reliability estimates across studies (Vacha-Haase, 1998). The results of the present study showed that RG researchers typically included both instrument characteristics and sample characteristics in their data analyses as variables that could influence reliability estimates. The most frequently coded instrument characteristics were response scale system and
instrument form and the most frequently coded sample characteristics were sample size, gender, age, and population type. Similarly, Vacha-Haase and Thompson (2011) found in their meta-analysis of RG studies that gender, sample size, age, and ethnicity were the most commonly reported sample characteristics.

**Optimal recommendations.** Romano and Kromrey (2002) argued that one issue occurring in some RG analyses was that the reliability estimates analyzed in the RG studies did not represent independent observations. Lack of independence of observations can include analyzing estimates from multiple subgroups, multiple types of reliability, multiple subscales, and/or multiple instruments. In the present study, typically researchers conducted separate analyses by subscale when they had multiple subscales, by instrument when they had multiple instruments, or by type of reliability estimate when they had multiple types of reliability estimates. Additionally, the results of the study by Dieckmann et al. (2009) showed that 66% of the meta-analyses in their study mentioned whether the study results were independent and gave details regarding how they treated dependencies if they occurred.

**Recommendations That Were not Followed**

The results showed that some RG recommendations were not typically followed by applied RG researchers.

**Essential recommendations.** The file-drawer problem is a problem facing any researcher conducting a meta-analytic study (Romano & Kromrey, 2002). The results of the present study show that many RG researchers did not check for publication bias or file-drawer bias, which is consistent the hypothesis that there would be a disconnect between the recommended practice of examining publication bias and the reporting of
publication bias by RG researchers. The results of the study by Dieckmann et al. (2009) showed that 5% of meta-analyses examined in their study used a funnel plot, 32% used a file-drawer analysis, and 4% of studies used both funnel plots and file-drawer analyses.

There are various ways that researchers can address the problem of missing data in RG studies. The results showed that there were eight RG studies in the present study and in the study by Vacha-Haase and Thompson (2011) that contacted authors to request reliability information. Another way to address the issue of missing data is to use the KR21 formula (Kuder & Richardson, 1937) when an instrument uses a dichotomous scale. Some authors in the present study used the KR21 formula to generate reliability information when reliability estimates were not given.

According to Dieckmann et al. (2009), unreliability in coding procedures of meta-analytic studies is a serious issue in that it can add random variation to the analysis. However, this issue can be dealt with by using multiple raters and calculating interrater reliability. The results of the present study showed that the majority of researchers did not calculate interrater reliability for the coding of their studies. The result of the study by Dieckmann et al. (2009) also demonstrated that 66% of the meta-analyses examined in their study did not report a method of coding interrater reliability.

One way to characterize variability is to use a box and whisker plot as recommended by Cousin and Henson (2000). However, the results showed that the majority of authors did not use a box and whisker plot to present their results. This is similar to the results found by Vacha-Haase and Thompson where only 35% of RG studies used a box and whisker plot. Researchers can also characterize variability with CIs; however, the results showed that researchers typically did not report CIs.
Dieckmann et al. (2009) noted that of the 91 studies that reported a measure of central tendency, 56% reported CIs.

**Optimal recommendations.** Four of the optimal recommendations for RG research were also frequently not followed. The results of the present study showed that a majority of researchers who included subgroups in their analyses combined multiple subgroups in one analysis without any clear justification for doing so. Similarly, most of the RG studies reviewed did not test for homogeneity of variance. Dieckmann et al. (2009) noted that of the 91 studies that reported some measure of variability, 59% reported a homogeneity test statistic. Additionally, the majority of authors did not conduct an a priori power analysis, which supports my hypothesis that many of the applied RG studies would not conduct an a priori power analysis. Similarly, there were not any meta-analyses in the study by Dieckmann et al. that conducted an a priori power analysis.

Although RG researchers did not follow some of the recommendations, it is important to note that it is possible that they did not do so because the journals they published in had space or other restrictions.

**Controversial Recommendations**

The three controversial recommendations examined in the present study concerned data transformation, weighting, and data analysis. The results of the current study showed that most authors did not transform the data or use a weighting approach. Additionally, the results showed that in general the number of studies that used a weighting approach increased after 2002. Furthermore, the results of the study by Dieckmann et al. (2009) showed that 71% of their sample used a weighting approach.
Additionally, researchers generally conducted fixed-effects analyses; this finding was similar to the results of the meta-analytic study by Vacha-Haase and Thompson (2011) wherein 81% of RG studies used multiple regression or ANOVA.

**Lack of Reporting of Reliability Estimates in Primary Studies**

As previously discussed in the introduction section, the perception that reliability is a property of instruments is problematic as it has led to an underreporting of reliability estimates and a widespread dismissal of its importance in research (Cousin & Henson, 2000). The trend of underreporting of reliability estimates was seen in the present study. In the present study, the RG studies that gave a breakdown of the reports of reliability estimates from primary studies were examined and the results showed that 55% of primary studies did not mention or report reliability estimates. Furthermore, only 24% of authors of primary studies reported the reliability estimates of their sample. Similarly, Vacha-Haase and Thompson (2011) found that 55% of the primary studies in their review of RG studies did not mention reliability. This result is discouraging because authors have been encouraged to report sample reliabilities for over a decade (see e.g., Wilkinson and the APA Task Force on Statistical Inference, 1999).

The lack of reporting of the sample reliability estimates also had practical implications for RG researchers. For example, some authors (17%) used a different data analytic technique than they wanted to employ because they were limited by the small number of reliability estimates reported in the primary studies. Additionally, seven RG researchers noted that they dropped a particular type of reliability estimate from the data analysis because there were a limited number of reliability estimates reported in the primary studies.
**Other RG Studies**

The RG study by Bornmann et al. (2010) was removed from all the main analyses because it examined interrater reliability and it did not look at sources of variability in reliability estimates. This study also did not examine a particular instrument or group of instruments that examined the same construct like every other RG study. However, it was a well-done study that followed many of the essential and optimal RG recommendations. Future researchers can examine this study if they are interested in conducting a RG study that examines interrater reliability estimates.

Dissertations and theses were not included in the main analyses because they have not been vetted in the same manner as other studies. However, the results showed that dissertations and theses were similar to other RG studies in that they followed most of the essential RG recommendations and some of the optimal recommendations.

**Exemplary RG Study**

Throughout the present paper, the current practices of RG researchers have been discussed, and these practices have been compared to the RG recommendations. For those readers who are interested in examining an exemplary RG study, I recommend the RG study conducted by Graham and Christiansen (2009). These researchers followed almost all of the essential recommendations. They used multiple keywords when searching for studies, used multiple methods of searching for studies, checked for publication bias, coded both instrument and sample characteristics, calculated interrater reliability, and reported CIs. They also followed multiple optimal recommendations. They conducted separate analyses of different subscales and instruments, and conducted a test for homogeneity of variance. In regards to the controversial recommendations,
Graham and Christiansen transformed the data before conducting the analyses, used a weighting approach, and conducted random-effects analyses. Additionally, they also provided a rationale for selecting their instruments, and provided details for how reliability was reported in the primary studies. This was a well-done study, and I would recommend that future RG researchers examine this study if they are looking for an example of an excellent RG study.

The current recommendations were written in a plethora of papers that span multiple academic journals and years. Therefore, it is not surprising that even this exemplary study did not follow all of the RG recommendations. For example, they combined multiple subgroups in one analysis, and did not conduct an a priori power analysis.

**Recommendations for Conducting RG Studies**

After reviewing many RG studies, our research team discussed different ways that RG studies could be improved. These recommendations can help future RG researchers conduct RG studies.

RG researchers should provide a rationale as to why they selected their instruments. This information will provide readers with the knowledge as to why a RG study was conducted with the instrument(s). The results showed that most authors did provide a rationale for selecting the RG instrument(s), such as the instrument being the most popular instrument used to measure the construct.

RG researchers should be clear when they report the number of reliability estimates that were used in their study. At times, it was difficult to determine the number of reliability estimates that were included in the data analysis. Additionally, it would be
helpful if RG researchers noted the number of primary studies that did not mention reliability, mentioned reliability but did not cite reliability values, or did reliability induction so that trends in the reporting of reliability estimates can be examined.

RG researchers should clearly state whether they used subgroups, whole groups, or both subgroups and whole groups and should provide a rationale for their choice. At times, it was challenging in the present study to determine whether subgroups were used (e.g., when an instrument included subscales, the additional estimates might have been due to subscales not subgroups). Additionally, multiple authors argued that the subgroups included in their studies were independent but did not explain why they thought this was the case. RG researchers should provide the reason why the subgroups were independent if they contend that they are independent.

RG researchers should be certain that all the information included in the tables matches what was written in the text. For example, if a variable is included in a table on data analyses, the variable should also be included in the discussion of variables that are coded.

It would be helpful if RG researchers clearly stated the inclusion criteria for studies, such as whether conference papers, dissertations, or papers published in other languages were included. Authors should be open to including additional sources other than journal papers (e.g., conference papers) so that a wider range of studies can be included. If authors do search for conference papers, it would be helpful to include this information in the text of the paper. It was possible that some RG studies searched for conference papers but did not find any, and so it seemed that they excluded conference papers from their search for studies.
As previously noted in the literature review section, Dieckmann et al. (2009) advised authors to determine whether sample size is confounded with other characteristics of the study sample before weighting by sample size. Researchers can run the analyses with and without using a weighting approach to see if the results differ.

**Limitations**

There are several potential limitations of this study. For example, this sample was limited because it did not include papers published after 2010. Furthermore, only papers that were written in English were included, which restricts the total number of papers. Additionally, the analyses were limited to those variables that were included in the coding. The focus of the present study was on the methods used; therefore, it did not examine the results that were found in RG studies. This study was also limited because it was an exploratory study; future studies can conduct additional analyses of RG studies.

**Importance of the Present Study**

Many differing recommendations exist in regard to how researchers should retrieve, code, and analyze information when conducting RG studies. The present study is significant because it was the first study that has done a comparison between RG recommendations and the practices of applied researchers.

The results of the present study showed that there were some differences between the RG recommendations and the practices of applied RG researchers (e.g., most RG researchers did not calculate interrater reliability). One potential reason why RG researchers may not follow RG recommendations is that the RG studies were influenced by the instrument(s) under review and the data available in the primary studies (e.g., some researchers wanted to include a particular type of reliability estimate in their
analyses but not enough reliability estimates were reported in the primary studies). RG researchers should be aware of the RG recommendations, but if a researcher has a valid reason for not adhering to a recommended practice, he or she should inform the reader why he or she deviated from it. Another reason RG researchers may not be following the RG recommendations is that many recommendations were published in specialized journals such as Educational and Psychological Measurement, Psychological Methods, or were discussed in measurement and statistics books.

Researchers have been conducting RG studies for over 10 years since it was first discussed by Vacha-Haase (1998). However, the number of studies published per year varied from one to 12. The results showed that there were only two RG studies published in 2010; however, six studies were published in 2009. The author searched the ERIC and PsycINFO databases using combinations of the key terms reliability, generalization and meta-analysis from January 01, 2011 to June 30, 2012 to examine the current trend in the publishing of RG papers. The results showed that nine RG studies were published in 2011 and zero in 2012 as of June 30, 2012. Therefore, there was not a clear trend in the number of RG papers being published in the past few years, but it seems that current researchers are still interested in using the RG technique. Many of the RG articles were published in Educational and Psychological Measurement, which applied researchers may not be reading.

**Future Directions**

According to Vacha-Haase and Thompson (2011), RG studies are important because the results of RG studies show that score reliabilities vary across administrations, which challenges the common misconception that instruments are reliable. The results of
RG studies may encourage future applied researchers to report the reliability of their sample.

RG studies can be informative if reliability data is missing. For example, the results of the RG study by Graham and Christiansen (2009) showed that Rubin’s (1970) Liking scale had a mean reliability of .887 across studies. If a researcher was examining a study of Rubin’s Liking scale that did not report the reliability estimates of the sample, the researcher could examine the results of the aforementioned RG study to see the average reliability estimate of this scale.

RG studies can also provide researchers with information regarding factors that influence reliability that they may want to take into account. For example, if a RG study found that instrument length influenced the reliability estimates, the researcher may choose to use the version of the instrument that was found to be more reliable.

Although the present study was focused on RG, it could also be useful for future researchers who are conducting other types of meta-analytic studies. For example, the present study provides recommendations for ways to search and code studies that could also be relevant for other meta-analytic researchers.

In summary, RG studies can be very useful for researchers when they are choosing instruments, so it is important that such studies be well designed. The present study indicates that there is still much room for improvement and it lays out a model for how to proceed.
Appendix A:

Applied RG Studies Included in the Data Analysis


Appendix B:

Code Sheet Used for Coding Studies

<table>
<thead>
<tr>
<th>Article Notes</th>
<th>0</th>
<th>Notes about articles</th>
<th>String</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Identifying Information:</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1a Who are the authors?</td>
<td>String</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1b What is the year of publication?</td>
<td>Numeric</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1c Numeric code for the paper</td>
<td>Numeric</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1d rater</td>
<td>1 = Allie 2 = Dr. T 3 = Angie</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rationale for Selecting an Instrument:</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2 Did the RG researchers give a rationale for selecting the instrument?</td>
<td>0 = no 1 = yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2a If yes was it because it was the most popular test (or commonly/widely used)</td>
<td>0 = no 1 = yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2b If yes was it because the author wanted to find which test had the best reliability among several tests</td>
<td>0 = no 1 = yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2c If yes was it because: author selected test(s) bc superior measure of a construct</td>
<td>0 = no 1 = yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2d Other</td>
<td>0 = no 1 = yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2dd If “other”, please specify.</td>
<td>String</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measures Coded:</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>3 How many measures did authors code?</td>
<td>Numeric</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Did the measure coded include subscales?</td>
<td>0 = no 1 = yes</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Article Search:</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Did the author of the RG study search for articles using:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5a Review articles</td>
<td>0 = no, 1 = yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5b ERIC</td>
<td>0 = no, 1 = yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5c PsycINFO or PsychLit</td>
<td>0 = no, 1 = yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5d References</td>
<td>0 = no, 1 = yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5e Contact authors for RG studies</td>
<td>0 = no, 1 = yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5f How many other sources were used</td>
<td>Numeric</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Article Search Keywords:

*What keyword(s) did the authors use to search for articles?*

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>6a</td>
<td>Title of measure</td>
<td>0 = no, 1 = yes</td>
</tr>
<tr>
<td>6b</td>
<td>Acronyms (and/or abbreviations)</td>
<td>0 = no, 1 = yes</td>
</tr>
<tr>
<td>6c</td>
<td>Construct</td>
<td>0 = no, 1 = yes</td>
</tr>
<tr>
<td>6d</td>
<td>Scale Author</td>
<td>0 = no, 1 = yes</td>
</tr>
<tr>
<td>6e</td>
<td>OTHER DISTINCT APPROACH, specify</td>
<td>String</td>
</tr>
<tr>
<td>6f</td>
<td>Notes on choosing reports</td>
<td>String</td>
</tr>
</tbody>
</table>

### Choosing Reports:

*When choosing studies to analyze in RG, were the following inclusion or exclusion criteria?*

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>7a</td>
<td>Conference/unpublished papers</td>
<td>0 = exclusion, 1 = inclusion, 2 = no mention</td>
</tr>
<tr>
<td>7b</td>
<td>Book chapters/books</td>
<td>0 = exclusion, 1 = inclusion, 2 = no mention</td>
</tr>
<tr>
<td>7c</td>
<td>Dissertations/theses</td>
<td>0 = exclusion, 1 = inclusion, 2 = no mention</td>
</tr>
<tr>
<td>7d</td>
<td>Journal articles/published papers</td>
<td>0 = exclusion, 1 = inclusion, 2 = no mention</td>
</tr>
<tr>
<td>7e</td>
<td>Paper written in language(s) other than English</td>
<td>0 = exclusion, 1 = inclusion, 2 = no mention</td>
</tr>
<tr>
<td>7f</td>
<td>Measure administered in a non-English language (translations!)</td>
<td>0 = exclusion, 1 = inclusion, 2 = no mention</td>
</tr>
</tbody>
</table>

### Starting Date:

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>Did the authors select a year as a starting date?</td>
<td>0 = no, 1 = yes</td>
</tr>
<tr>
<td>8a</td>
<td>If authors selected a year, what reason was given for selecting this date?</td>
<td>1 = 1st yr of test publication (stated) 2 = yr new test form was published 3 = to make RG study manageable 4 = none given 5 = 1st yr of test publication (not stated) 6 = other</td>
</tr>
<tr>
<td>8aa</td>
<td>if other reasons were given, specify</td>
<td>String</td>
</tr>
</tbody>
</table>
### Coding Test Characteristics:
*When coding studies for RG analysis, was this test characteristic coded for?*

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
<th>Coding Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>9a</td>
<td>Did response scale type vary across primary studies included in RG?</td>
<td>0 = no, 1 = yes</td>
</tr>
<tr>
<td>9aa</td>
<td>Type of response scale</td>
<td>0 = not coded, 1 = coded</td>
</tr>
<tr>
<td>9aaa</td>
<td>How was type of response scale coded?</td>
<td>String</td>
</tr>
<tr>
<td>9aaaa</td>
<td>If response scale was coded, was it analyzed</td>
<td>0 = not used in analysis, 1 = used in analysis</td>
</tr>
<tr>
<td>9b</td>
<td>Did “who completed the measure” vary across primary studies included in RG? (i.e. student vs. parent)?</td>
<td>0 = no, 1 = yes</td>
</tr>
<tr>
<td>9bb</td>
<td>Who completed the measure</td>
<td>0 = not coded, 1 = coded</td>
</tr>
<tr>
<td>9bbb</td>
<td>How was “who completed the measure” coded?</td>
<td>String</td>
</tr>
<tr>
<td>9bbbb</td>
<td>If “who completed measure” was coded, was it analyzed</td>
<td>0 = not used in analysis, 1 = used in analysis</td>
</tr>
<tr>
<td>9c</td>
<td>Were multiple translations (languages) of the instrument included in the RG study?</td>
<td>0 = no, 1 = yes</td>
</tr>
<tr>
<td>9cc</td>
<td>Language of the test</td>
<td>0 = not coded, 1 = coded</td>
</tr>
<tr>
<td>9ccc</td>
<td>If multiple translations were coded, was it analyzed</td>
<td>0 = not used in analysis, 1 = used in analysis</td>
</tr>
<tr>
<td>9d</td>
<td>Were multiple forms of any individual measure included in the RG study?</td>
<td>0 = no, 1 = yes</td>
</tr>
<tr>
<td>9dd</td>
<td>Form of test used</td>
<td>0 = not coded, 1 = coded</td>
</tr>
<tr>
<td>9ddd</td>
<td>If multiple forms were coded, was it analyzed</td>
<td>0 = not used in analysis, 1 = used in analysis</td>
</tr>
</tbody>
</table>

### Coding Sample Characteristics:
*Were the following coded for in this particular RG study?*

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
<th>Coding Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>10a</td>
<td>Sample Size</td>
<td>0 = no, 1 = yes</td>
</tr>
<tr>
<td>10aa</td>
<td>If sample size was coded, was it analyzed</td>
<td>0 = not used in analysis, 1 = used in analysis</td>
</tr>
<tr>
<td>10aaa</td>
<td>Notes on Sample Size</td>
<td>String</td>
</tr>
<tr>
<td>10b</td>
<td>Gender homogeneity/proportion</td>
<td>0 = no, 1 = yes</td>
</tr>
<tr>
<td>10bb</td>
<td>If gender was coded, how?</td>
<td>String</td>
</tr>
<tr>
<td>10bbb</td>
<td>If gender was coded, was it analyzed</td>
<td>0 = not used in analysis, 1 = used in analysis</td>
</tr>
<tr>
<td>10c</td>
<td>Racial/ethnic homogeneity/proportion</td>
<td>0 = no, 1 = yes</td>
</tr>
<tr>
<td>10cc</td>
<td>If race/ethnicity coded, how?</td>
<td>String</td>
</tr>
<tr>
<td>Column</td>
<td>Description</td>
<td>Notes</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
<td>-------</td>
</tr>
<tr>
<td>10ccc.</td>
<td>If race/ethnicity was coded, was it analyzed</td>
<td>0 = not used in analysis, 1 = used in analysis</td>
</tr>
<tr>
<td>10cccc</td>
<td>Notes on Race/Ethnicity</td>
<td>String</td>
</tr>
<tr>
<td>10d</td>
<td>Age</td>
<td>0 = no, 1 = yes</td>
</tr>
<tr>
<td>10dd</td>
<td>If age was coded, how?</td>
<td>0 = continuous variable, 1 = group variable, 2 = mean age</td>
</tr>
<tr>
<td>10dddd</td>
<td>If age was grouped, how?</td>
<td>String</td>
</tr>
<tr>
<td>10dddd</td>
<td>If age was coded, was it analyzed</td>
<td>0 = not used in analysis, 1 = used in analysis</td>
</tr>
<tr>
<td>10e</td>
<td>Population type</td>
<td>0 = no, 1 = yes</td>
</tr>
<tr>
<td>10ee</td>
<td>If pop type coded, how?</td>
<td>String</td>
</tr>
<tr>
<td>10eee</td>
<td>If population type was coded, was it analyzed</td>
<td>0 = not used in analysis, 1 = used in analysis</td>
</tr>
<tr>
<td>10eeee</td>
<td>Population Notes</td>
<td>String</td>
</tr>
<tr>
<td>10f</td>
<td>What other sample characteristics were coded?</td>
<td>String</td>
</tr>
<tr>
<td>10ff</td>
<td>If other characteristics were coded, were they analyzed (note each variable)</td>
<td>0 = not used in analysis, 1 = used in analysis</td>
</tr>
<tr>
<td>10g</td>
<td>List characteristics the author wanted to code for but couldn’t</td>
<td>String</td>
</tr>
</tbody>
</table>

**RG researchers Inter-Rater Reliability for Coding:**

<table>
<thead>
<tr>
<th>Column</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>Did the authors do IRR for coding?</td>
</tr>
<tr>
<td>11a</td>
<td>IF IRR checked, how many raters were used?</td>
</tr>
<tr>
<td>11aa</td>
<td>IRR notes</td>
</tr>
</tbody>
</table>

**Methods of Computing Author’s IRR: If IRR was computed, was it computed using:**

<table>
<thead>
<tr>
<th>Column</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>12a</td>
<td>Percent agree</td>
</tr>
<tr>
<td>12b</td>
<td>Cohen’s kappa</td>
</tr>
<tr>
<td>12c</td>
<td>Intra-Class Correlations</td>
</tr>
<tr>
<td>12d</td>
<td>Not given</td>
</tr>
</tbody>
</table>

**Reliability Used in Actual RG Data Analysis:**

<table>
<thead>
<tr>
<th>Column</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>13a</td>
<td>IRR (inter-rater reliability)</td>
</tr>
<tr>
<td>13b</td>
<td>KR-20</td>
</tr>
<tr>
<td>13c</td>
<td>Alpha</td>
</tr>
<tr>
<td>13d</td>
<td>Split-half</td>
</tr>
<tr>
<td>13e</td>
<td>Test-Retest</td>
</tr>
<tr>
<td>13f</td>
<td>Parallel Forms</td>
</tr>
<tr>
<td>13g</td>
<td>if applicable, list any types of rel. that were dropped from the more sophisticated analyses - and why</td>
</tr>
</tbody>
</table>
**Coding Aspects of the Test Instrument:**

*Were the following coded for?*

<table>
<thead>
<tr>
<th>14a</th>
<th>Were measures (or versions) of varying length included?</th>
<th>0 = no, 1 = yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>14aa</td>
<td>If measure/version length varied, was test length coded for?</td>
<td>0 = not coded for, 1 = coded for</td>
</tr>
<tr>
<td>14aaa</td>
<td>If measure/version length varied, was it analyzed</td>
<td>0 = not used in analysis, 1 = used in analysis</td>
</tr>
<tr>
<td>14b</td>
<td>Correction for attenuation</td>
<td>0 = not coded for, 1 = coded for</td>
</tr>
<tr>
<td>14c</td>
<td>Did they code SD or variance</td>
<td>0 = not used for, 1 = coded for</td>
</tr>
<tr>
<td>14cc</td>
<td>Did they use SD or variance as a predictor</td>
<td>0 = not used in analysis, 1 = used in analysis</td>
</tr>
<tr>
<td>14ccc</td>
<td>Did they use a weighting approach</td>
<td>0 = not used, 1 = used</td>
</tr>
<tr>
<td>14cccc</td>
<td>What type of weighting did they use (specify: N, SD, both or other)</td>
<td>String</td>
</tr>
<tr>
<td>14d</td>
<td>Did the authors report the reliability of:</td>
<td>1 = ONLY subgroups, 2 = ONLY whole groups, 3 = both subgroups and whole groups</td>
</tr>
</tbody>
</table>

**Independence Issues:**

| 15 | Were all reliability reports independent (pre-analysis)? | 0 = no, 1 = yes |

**Reasons for Non-independence: Multiple SUBSCALES:**

<table>
<thead>
<tr>
<th>15a</th>
<th>Reported reliability for more than one subscale</th>
<th>0 = no, 1 = yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>15aa</td>
<td>If multiple subscale reliabilities were reported, how was this handled during data analysis?</td>
<td>1 = conducted separate analyses (specify), 2 = combined multiple subscale reliability in one analysis, 3 = used only one type of subscale, 4 = used fixed-effect MANOVA, 5 = HLM, 6 = dummy coded</td>
</tr>
<tr>
<td></td>
<td>Description</td>
<td>Options</td>
</tr>
<tr>
<td>------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------</td>
</tr>
<tr>
<td>15aaa</td>
<td>If multiple subscales was 1, specify</td>
<td>String</td>
</tr>
<tr>
<td>15aaaa</td>
<td>If 2 is selected, - did they justify it (8) or not address the issue (7)</td>
<td>7 = issue not addressed (verbal or analysis)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8 = issue verbally addressed, not addressed in analysis</td>
</tr>
<tr>
<td>15aaaaa</td>
<td>If 8 in 15aaaa – what is the reason</td>
<td>String</td>
</tr>
</tbody>
</table>

**Reasons for Non-independence: Multiple SUBGROUPS:**

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
<th>Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>15b</td>
<td>Reported reliability for more than one subgroup</td>
<td>0 = no</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = yes</td>
</tr>
<tr>
<td>15bb</td>
<td>If multiple subgroup reliabilities were reported, how was this handled during data analysis?</td>
<td>1 = conducted separate analyses (specify)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = combined multiple subgroup reliabilities in one analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = used only one type of subgroup</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 = used fixed-effect MANOVA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5 = HLM</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6 = dummy coded</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7 = issue not addressed (verbal or analysis)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8 = issue verbally addressed, not addressed in analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9 = other (specify)</td>
</tr>
<tr>
<td>15bbbb</td>
<td>If 2 is selected in 15bb, - did they justify it (8) or not mention (7)</td>
<td>7 = issue not addressed (verbal or analysis)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8 = issue verbally addressed, not addressed in analysis</td>
</tr>
<tr>
<td>15bbbb</td>
<td>If 8 in 15 bbbb - what is the reason</td>
<td>String</td>
</tr>
</tbody>
</table>

**Reasons for Non-independence: Multiple MEASURES:**

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
<th>Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>15c</td>
<td>Reported reliability for more than one measure</td>
<td>0 = no</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = yes</td>
</tr>
<tr>
<td>15cc</td>
<td>If multiple measure reliabilities were reported, how was this handled during data analysis?</td>
<td>1 = conducted separate analyses (specify)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = combined multiple measure reliabilities in one analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = used only one type of measure</td>
</tr>
<tr>
<td>15ccc</td>
<td>If separate analysis for measure, specify</td>
<td>String</td>
</tr>
</tbody>
</table>
| 15ccc | If 2 is selected, - did they justify it (8) or not address the issue (7) | 7 = issue not addressed (verbal or analysis)  
8 = issue verbally addressed, not addressed in analysis |
| 15cccc | If 8 in 15ccc what is the reason | String |

**Reasons for Non-independence: Multiple RELIABILITY TYPES (i.e. alpha and test-retest):**

| 15d | Reported multiple types of reliabilities | 0 = no  
1 = yes |
| 15dd | If multiple types of reliabilities were reported, how was this handled during data analysis? | 1 = conducted separate analyses (specify)  
2 = combined multiple types of reliabilities in one analysis  
3 = used only one type of reliability  
4 = used fixed-effect MANOVA  
5 = HLM  
6 = dummy coded  
7 = issue not addressed (verbal or analysis)  
8 = issue verbally addressed, not addressed in analysis  
9 = other (specify)  
10 = 1 and 2 |
| 15ddd | If separate analysis (1 or 10) for multiple types, specify | String |
| 15dddd | If 2 or 10 (1 & 2) is selected, - did they justify it (8) or not address the issue (7) | 7 = issue not addressed (verbal or analysis)  
8 = issue verbally addressed, not addressed in analysis |
| 15dddd | If 8 in 15ddddd what is the reason | String |
### Publication Bias:

| 16 | Did the authors check for publication bias or file drawer bias? | 0 = no  
 |    | 1 = yes |  |
| 16a | If this was checked, what method was used? | 0 = method not given  
 |    | 1 = Begg’s test  
 |    | 2 = funnel plot and failsafe n  
 |    | 3 = other (specify)  |
| 16aa | Additional details of how it was checked | String |

### Missing Data:

| 17 | Were studies included that did not report reliability? | 0 = no, 1 = yes |
| 17a | Did the measure(s) use a dichotomous scale | 0 = no, 1 = yes |
| 17b | Did the author’s use KR-21? | 0 = no, 1 = yes |
| 17c | Did author’s email article author’s for reliability information? | 0 = no, 1 = yes |
| 17d | If other methods were used, specify. | String |

### Transformations:

| 18 | Did authors transform data for analysis? | 0 = no, 1 = yes |

### Data used for Analysis:

**Did the RG researchers use for the reliability coefficients . . . (more than one can apply)**

| 18.5a | Raw data | 0 = not used, 1 = used |
| 18.5b | Square root of reliability | 0 = not used, 1 = used |
| 18.5c | Fisher’s r-to-z | 0 = not used, 1 = used |
| 18.5d | Squared reliability coefficient | 0 = not used, 1 = used |
| 18.5e | If a different method was used, specify | String |

### Power:

| 19 | Did authors conduct an a priori power analysis? | 0 = no, 1 = yes |
| 19a | Power analysis notes | String |

### Homogeneity of Variances:

| 20 | Did authors conduct a test for homogeneity of population correlations? | 0 = no, 1 = yes |
**Homogeneity was tested with . . .**

<table>
<thead>
<tr>
<th></th>
<th>HLM random effects test</th>
<th>0 = no, 1= yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>20a</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Q test</th>
<th>0 = no, 1= yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>20b</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>A different method</th>
<th>0 = no, 1= yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>20c</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>If a different method was used, specify.</th>
<th>String</th>
</tr>
</thead>
<tbody>
<tr>
<td>20d</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Data Analysis:**

<table>
<thead>
<tr>
<th></th>
<th>Type of Data Analysis</th>
<th>0 = fixed effects 1 = random effects 2 = HLM 3 = descriptive (primary analysis) 4 = correlations (primary analysis) 5 = other (specify)</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>If other, specify</th>
<th>String</th>
</tr>
</thead>
<tbody>
<tr>
<td>21a</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Data Analysis Justification:**

<table>
<thead>
<tr>
<th></th>
<th>Did researchers justify the type of data analysis?</th>
<th>0 = no, 1= yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>If justified, specify.</th>
<th>String</th>
</tr>
</thead>
<tbody>
<tr>
<td>22a</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Confidence Intervals:**

<table>
<thead>
<tr>
<th></th>
<th>Were CI reported?</th>
<th>0 = no, 1= yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>If reported, what were the CIs based upon?</th>
<th>1 = fixed model 2 = random/mixed effects model</th>
</tr>
</thead>
<tbody>
<tr>
<td>23a</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Number of Articles:**

<table>
<thead>
<tr>
<th></th>
<th># of primary studies w/ reported reliability estimates found by RG author</th>
<th>Numeric</th>
</tr>
</thead>
<tbody>
<tr>
<td>24</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Notes on number of primary studies</th>
<th>String</th>
</tr>
</thead>
<tbody>
<tr>
<td>24a</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Total number of reliability estimates used in RG analysis</th>
<th>String</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Reliability: Measure 1 Internal Consistency**

<table>
<thead>
<tr>
<th></th>
<th>Number of IC reliability estimates for Measure 1</th>
<th>Numeric</th>
</tr>
</thead>
<tbody>
<tr>
<td>26</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>If Measure 1 IC had multiple types of IC - how many were in each type</th>
<th>Numeric</th>
</tr>
</thead>
<tbody>
<tr>
<td>26.25</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Notes on the Measure 1 IC reliability estimates</th>
<th>String</th>
</tr>
</thead>
<tbody>
<tr>
<td>26.50</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Number of IC reliability estimates for Measure 1 Subscale 1</th>
<th>Numeric</th>
</tr>
</thead>
<tbody>
<tr>
<td>26a</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Number of IC reliability estimates for Measure 1 Subscale 2</th>
<th>Numeric</th>
</tr>
</thead>
<tbody>
<tr>
<td>26b</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26c</td>
<td>Number of IC reliability estimates for Measure 1 Subscale 3</td>
<td>Numeric</td>
</tr>
<tr>
<td>26d</td>
<td>Number of IC reliability estimates for Measure 1 Subscale 4</td>
<td>Numeric</td>
</tr>
<tr>
<td>26e</td>
<td>Number of IC reliability estimates for Measure 1 Subscale 5</td>
<td>Numeric</td>
</tr>
<tr>
<td>26f</td>
<td>Number of IC reliability estimates for Measure 1 Subscale 6</td>
<td>Numeric</td>
</tr>
<tr>
<td>26g</td>
<td>Number of IC reliability estimates for Measure 1 Subscale 7</td>
<td>Numeric</td>
</tr>
<tr>
<td>26h</td>
<td>Number of IC reliability estimates for Measure 1 Subscale 8</td>
<td>Numeric</td>
</tr>
<tr>
<td>26i</td>
<td>Number of IC reliability estimates for Measure 1 Subscale 9</td>
<td>Numeric</td>
</tr>
<tr>
<td>26j</td>
<td>Number of IC reliability estimates for Measure 1 Subscale 10</td>
<td>Numeric</td>
</tr>
<tr>
<td>26k</td>
<td>Number of IC reliability estimates for Measure 1 Subscale 11</td>
<td>Numeric</td>
</tr>
</tbody>
</table>

**Reliability: Measure 1 Test-Retest**

| 27  | Number of T-RT reliability estimates for Measure 1 | Numeric |
| 27.5| Notes on Measure 1 T-RT | String |
| 27a | Number of T-RT reliability estimates for Measure 1 Subscale 1 | Numeric |
| 27b | Number of T-RT reliability estimates for Measure 1 Subscale 2 | Numeric |
| 27c | Number of T-RT reliability estimates for Measure 1 Subscale 3 | Numeric |
| 27d | Number of T-RT reliability estimates for Measure 1 Subscale 4 | Numeric |
| 27e | Number of T-RT reliability estimates for Measure 1 Subscale 5 | Numeric |
| 27f | Number of T-RT reliability estimates for Measure 1 Subscale 6 | Numeric |
| 27g | Number of T-RT reliability estimates for Measure 1 Subscale 7 | Numeric |
| 27h | Number of T-RT reliability estimates for Measure 1 Subscale 8 | Numeric |
| 27i | Number of T-RT reliability estimates for Measure 1 Subscale 9 | Numeric |
| 27j | Number of T-RT reliability estimates for Measure 1 Subscale 10 | Numeric |

**Reliability: Measure 2 Internal Consistency**

| 28  | Number of IC reliability estimates for Measure 2 | Numeric |
| 28.5| Notes on measure 2 IC | String |
### Reliability: Measure 2 Test-Retest

<table>
<thead>
<tr>
<th></th>
<th>Number of T-RT reliability estimates for Measure 2</th>
<th>Numeric</th>
</tr>
</thead>
<tbody>
<tr>
<td>29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29a</td>
<td>Number of T-RT reliability estimates for Measure 2</td>
<td></td>
</tr>
<tr>
<td>29b</td>
<td>Subscale 1</td>
<td></td>
</tr>
<tr>
<td>29c</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29d</td>
<td>Number of T-RT reliability estimates for Measure 2</td>
<td></td>
</tr>
<tr>
<td>29e</td>
<td>Subscale 2</td>
<td></td>
</tr>
<tr>
<td>29f</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29g</td>
<td>Number of T-RT reliability estimates for Measure 2</td>
<td></td>
</tr>
<tr>
<td>29h</td>
<td>Subscale 3</td>
<td></td>
</tr>
<tr>
<td>29i</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29j</td>
<td>Number of T-RT reliability estimates for Measure 2</td>
<td></td>
</tr>
<tr>
<td>29k</td>
<td>Subscale 4</td>
<td></td>
</tr>
</tbody>
</table>

### Reliability: Measure 3 Internal Consistency

<table>
<thead>
<tr>
<th></th>
<th>Number of IC reliability estimates for Measure 3</th>
<th>Numeric</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30.5</td>
<td>Notes on Measure 3 IC</td>
<td>String</td>
</tr>
<tr>
<td>30a</td>
<td>Number of IC reliability estimates for Measure 3</td>
<td></td>
</tr>
<tr>
<td>30b</td>
<td>Subscale 1</td>
<td></td>
</tr>
<tr>
<td>30c</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30d</td>
<td>Number of IC reliability estimates for Measure 3</td>
<td></td>
</tr>
<tr>
<td>30e</td>
<td>Subscale 2</td>
<td></td>
</tr>
<tr>
<td>30f</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30g</td>
<td>Number of IC reliability estimates for Measure 3</td>
<td></td>
</tr>
<tr>
<td>30h</td>
<td>Subscale 3</td>
<td></td>
</tr>
<tr>
<td>30i</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30j</td>
<td>Number of IC reliability estimates for Measure 3</td>
<td></td>
</tr>
<tr>
<td>30k</td>
<td>Subscale 4</td>
<td></td>
</tr>
</tbody>
</table>

### Reliability: Measure 3 Test-Retest

<table>
<thead>
<tr>
<th></th>
<th>Number of T-RT reliability estimates for Measure 3</th>
<th>Numeric</th>
</tr>
</thead>
<tbody>
<tr>
<td>31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31a</td>
<td>Number of T-RT reliability estimates for Measure 3</td>
<td></td>
</tr>
<tr>
<td>31b</td>
<td>Subscale 1</td>
<td></td>
</tr>
<tr>
<td>31c</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>31c</td>
<td>Number of T-RT reliability estimates for Measure 3 Subscale 3</td>
<td></td>
</tr>
<tr>
<td>31d</td>
<td>Number of T-RT reliability estimates for Measure 3 Subscale 4</td>
<td></td>
</tr>
</tbody>
</table>

**Reliability: Measure 4 Internal Consistency**

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>Number of IC reliability estimates for Measure 4</td>
<td></td>
<td>Numeric</td>
</tr>
<tr>
<td>32a</td>
<td>Number of IC reliability estimates for Measure 4 Subscale 1</td>
<td></td>
<td>Numeric</td>
</tr>
<tr>
<td>32b</td>
<td>Number of IC reliability estimates for Measure 4 Subscale 2</td>
<td></td>
<td>Numeric</td>
</tr>
<tr>
<td>32c</td>
<td>Number of IC reliability estimates for Measure 4 Subscale 3</td>
<td></td>
<td>Numeric</td>
</tr>
<tr>
<td>32d</td>
<td>Number of IC reliability estimates for Measure 4 Subscale 4</td>
<td></td>
<td>Numeric</td>
</tr>
</tbody>
</table>

**Reliability: Measure 4 Test-Retest**

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>33</td>
<td>Number of T-RT reliability estimates for Measure 4</td>
<td></td>
<td>Numeric</td>
</tr>
<tr>
<td>33a</td>
<td>Number of T-RT reliability estimates for Measure 4 Subscale 1</td>
<td></td>
<td>Numeric</td>
</tr>
<tr>
<td>33b</td>
<td>Number of T-RT reliability estimates for Measure 4 Subscale 2</td>
<td></td>
<td>Numeric</td>
</tr>
<tr>
<td>33c</td>
<td>Number of T-RT reliability estimates for Measure 4 Subscale 3</td>
<td></td>
<td>Numeric</td>
</tr>
<tr>
<td>33d</td>
<td>Number of T-RT reliability estimates for Measure 4 Subscale 4</td>
<td></td>
<td>Numeric</td>
</tr>
</tbody>
</table>

**Reliability: Measure 5 Internal Consistency**

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>34</td>
<td>Number of IC reliability estimates for Measure 5</td>
<td></td>
<td>Numeric</td>
</tr>
<tr>
<td>34a</td>
<td>Number of IC reliability estimates for Measure 5 Subscale 1</td>
<td></td>
<td>Numeric</td>
</tr>
<tr>
<td>34b</td>
<td>Number of IC reliability estimates for Measure 5 Subscale 2</td>
<td></td>
<td>Numeric</td>
</tr>
<tr>
<td>34c</td>
<td>Number of IC reliability estimates for Measure 5 Subscale 3</td>
<td></td>
<td>Numeric</td>
</tr>
<tr>
<td>34d</td>
<td>Number of IC reliability estimates for Measure 5 Subscale 4</td>
<td></td>
<td>Numeric</td>
</tr>
</tbody>
</table>

**Reliability: Measure 5 Test-Retest**

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>35</td>
<td>Number of T-RT reliability estimates for Measure 5</td>
<td></td>
<td>Numeric</td>
</tr>
<tr>
<td>35a</td>
<td>Number of T-RT reliability estimates for Measure 5 Subscale 1</td>
<td></td>
<td>Numeric</td>
</tr>
<tr>
<td>35b</td>
<td>Number of T-RT reliability estimates for Measure 5 Subscale 2</td>
<td></td>
<td>Numeric</td>
</tr>
<tr>
<td>35c</td>
<td>Number of T-RT reliability estimates for Measure 5 Subscale 3</td>
<td></td>
<td>Numeric</td>
</tr>
<tr>
<td></td>
<td>Number of T-RT reliability estimates for Measure 5 Subscale 4</td>
<td>Numeric</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>----------------------------------------------------------</td>
<td>---------</td>
<td></td>
</tr>
</tbody>
</table>

**Reliability: measure 5 Inter-rater reliability**

<table>
<thead>
<tr>
<th>35.5</th>
<th>Number of IRR reliability estimates for Measure 5</th>
<th>Numeric</th>
</tr>
</thead>
</table>

* Note: One study had a total of 22 measures - therefore the coding spreadsheet includes variables 35.01-35.17 to have a column for each measure, but for space reasons each of these variables is not included in a table here

**Moderators:**

<table>
<thead>
<tr>
<th>36</th>
<th>Were predictors tested for in DA?</th>
<th>0 = no, 1= yes</th>
</tr>
</thead>
</table>
| 36a | If tested, how?                   | 1 = test pred. all at once  
2 = test pred. separately  
3 = test pred. in a series of blocks or hierarchical regression methods  
4 = other (specify)  
5 = tested but not reported how |
| 36aa| Moderator notes                   | String        |

**Variability in Reliability Coefficients:**

*To show variability in the reliability coefficients . . .*

<table>
<thead>
<tr>
<th>37a</th>
<th>Was a box and whisker plot used?</th>
<th>0 = no, 1= yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>37b</td>
<td>Were means and SD used?</td>
<td>0 = no, 1= yes</td>
</tr>
</tbody>
</table>

**Reporting of Reliability in primary studies as reported by RG researchers:**

*Note: report the number if it is given, if not report the percent*

<table>
<thead>
<tr>
<th>38</th>
<th>Did the RG researchers give the reports of reliability from the primary studies</th>
<th>0 = no, 1= yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>38a</td>
<td>Number of primary studies that did not mention reliability or did not report reliability</td>
<td>Numeric</td>
</tr>
<tr>
<td>38b</td>
<td>Number of primary studies that mentioned reliability but did not cite any values</td>
<td>Numeric</td>
</tr>
<tr>
<td>38c</td>
<td>Number of primary studies that did reliability induction</td>
<td>Numeric</td>
</tr>
<tr>
<td>38d</td>
<td>Number of primary studies that reported the reliability for their sample</td>
<td>Numeric</td>
</tr>
<tr>
<td>38e</td>
<td>If another category of reports</td>
<td>String</td>
</tr>
</tbody>
</table>
of the reliability from the primary studies was given, please describe the category and give the number in the category

| 38f | Notes on reporting reliability | String |

* Note: If a variable does not provide information, or information was not given when it should have been provided, the code 888 is used; for example if the RG researchers said they coded for age but did not specify how age was coded
* Note: If something is not applicable it is coded as 999
Appendix C:

Complete Codebook

TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>BACKGROUND INFORMATION</td>
<td>108</td>
</tr>
<tr>
<td>Rationale, Measures coded</td>
<td></td>
</tr>
<tr>
<td>METHODS: HOW RG RESEARCHERS GATHERED THE ARTICLES</td>
<td>109</td>
</tr>
<tr>
<td>Article search, Article search keywords, Choosing Reports, Starting dates</td>
<td></td>
</tr>
<tr>
<td>RG CHOICES: CODING THE TEST AND THE SAMPLE</td>
<td>113</td>
</tr>
<tr>
<td>Coding test characteristics, Coding sample characteristics,</td>
<td></td>
</tr>
<tr>
<td>INTRERRATER RELIABILITY</td>
<td>117</td>
</tr>
<tr>
<td>RG researchers IRR for coding, Computing IRR</td>
<td></td>
</tr>
<tr>
<td>INDEPENDENT AND DEPENDENT VARIABLES</td>
<td>118</td>
</tr>
<tr>
<td>Reliability used in analysis, Coding factors that affect reliability</td>
<td></td>
</tr>
<tr>
<td>ADDRESSING RECOGNIZED PROBLEMS: INDEPENDENCE</td>
<td>121</td>
</tr>
<tr>
<td>Reliability independence issues</td>
<td></td>
</tr>
<tr>
<td>ADDRESSING RECOGNIZED PROBLEMS: MISSING DATA</td>
<td>124</td>
</tr>
<tr>
<td>File drawer bias, Dealing with missing data</td>
<td></td>
</tr>
<tr>
<td>TRANSFORMATIONS, DATA USED FOR ANALYSIS, POWER, AND HOMOGENEITY OF VARIANCES</td>
<td>126</td>
</tr>
<tr>
<td>DATA ANALYSIS, CIs, MODERATORS, AND VARIANCE</td>
<td>127</td>
</tr>
</tbody>
</table>
**Background Information:**

**Article Notes**
0. Notes about articles
   *Variable name: notes*
   *Variable Type: string*

**Identifying Information:**
1a. Who are the authors?
   *Variable Name: author*
   *Variable Type: String*

1b. What is the year of publication?
   *Variable Name: year*
   *Variable Type: Numeric*

1c. Numeric code for the paper
   *Variable Name: numeric_code*
   *Variable Type: Numeric*

1d. rater
   *Variable Name: rater*
   *Variable Type: Numeric*
   **Codes:**
   1 = Allie
   2 = Dr. Toland
   3 = Angie

**Selecting an Instrument Rationale:**
2. Did the RG researchers give a rationale for selecting the analyzed instrument?
   *Variable Name: rationale*
   *Variable Type: Nominal*
   **Codes:**
   0 = no
   1 = yes
   *(If the answer is NO, skip to question 3)*

2a. If yes was it because it was the most popular test (or commonly/widely used)?
   *Variable Name: rat_given_popular*
   *Variable Type: Nominal*
   **Codes:**
   0 = no
   1 = yes

2b. If yes was it because it was the authors wanted to find which test had best reliability among several tests?
   *Variable Name: rat_given_best*
   *Variable Type: Nominal*
   **Codes:**
   0 = no
   1 = yes
2c. If yes was it because the authors selected the test(s) based on it/them being superior to other measure(s) of the construct?
   Variable Name: rat_given_superior
   Variable Type: Nominal
   Codes: 0 = no
           1 = yes

2d. If yes, was there another reason given?
   Variable Name: rat_given_other
   Variable Type: Nominal
   Codes: 0 = no
           1 = yes
   (If response was yes, please proceed to 2dd, otherwise skip to 3)

2dd. If the rationale given was “other”, please specify.
   Variable Name: rat_other_sp
   Variable Type: String

Measures Coded:
3. How many measures did authors code?
   Variable Name: num_measures
   Variable Type: Numeric

4. Did measure coded include subscales (just the presence of a subscale, not necessarily coded)?
   Variable Name: sub_measures
   Variable Type: Nominal
   Codes: 0 = no
           1 = yes

Methods: How RG Researchers Gathered Articles

Article Search:
5a. Did the author of the primary study search for articles using: review articles?
   Variable Name: review
   Variable Type: Nominal
   Codes: 0 = no
           1 = yes

5b. Did the author of the primary study search for articles using: ERIC?
   Variable Name: eric
   Variable Type: Nominal
   Codes: 0 = no
           1 = yes
5c. Did the author of the primary study search for articles using: PsycINFO or Psychlit?
   Variable Name: psychinfo
   Variable Type: Nominal
   Code:  0 = no
          1 = yes

5d. Did the author of the primary study search for articles using: References?
   Variable Name: references
   Variable Type: Nominal
   Codes:  0 = no
          1 = yes

5e. Did the author of the primary study search for articles by contacting authors for RG studies?
   Variable Name: contact_authors
   Variable Type: Nominal
   Codes:  0 = no
          1 = yes

5f. How many other sources were used?
   Variable Name: artsearch_other
   Variable Type: Numeric
   (If there are no other sources, skip to question 6, otherwise proceed to 5gg)

5g. If other sources were used, please specify.
   Variable Name: artsearch_othersp
   Variable Type: String

Article Search Keywords:
What keyword(s) did authors use to search for articles?
6a. Keywords used to search: title of measure
   Variable Name: title
   Variable Type: Nominal
   Code:  0 = not used
          1 = used

6b. Keywords used to search: Acronyms
   Variable Name: acronym
   Variable Type: Nominal
   Code:  0 = not used
          1 = used
6c. Keywords used to search: Construct
   Variable Name: construct
   Variable Type: Nominal
   Code: 0 = not used
         1 = used

6d. Keywords used to search: Scale author
   Variable Name: scaleauthor
   Variable Type: Nominal
   Code: 0 = not used
         1 = used

6e. Keywords used to search: other approach - specify
   Variable Name: search_other
   Variable Type: String

6f. Notes on choosing reports
   Variable Name: notes_choosing
   Variable Type: String

Choosing Reports:
7a. Choosing Reports: conference papers/unpublished
   Variable Name: conference_unpublished
   Variable Type: Nominal
   Code: 0 = exclusion criterion
         1 = inclusion criteria
         2 = no mention

7b. Choosing Reports: book chapters/books
   Variable Name: books
   Variable Type: Nominal
   Code: 0 = exclusion criterion
         1 = inclusion criteria
         2 = no mention

7c. Choosing Reports: dissertations/thesis
   Variable Name: dis_thesis
   Variable Type: Nominal
   Code: 0 = exclusion criterion
         1 = inclusion criteria
         2 = no mention
7d. Choosing Reports: journal articles/published papers

Variable Name: journal_published
Variable Type: Nominal

Code:
0 = exclusion criterion
1 = inclusion criteria

7e. Choosing reports: Paper written in language(s) other than English

Variable Name: language_paper
Variable Type: Nominal

Code:
0 = exclusion criterion
1 = inclusion criteria
2 = no mention

7f. Choosing Reports: measure administered in a non-English language (translations)

Variable Name: language
Variable Type: Nominal

Code:
0 = exclusion criterion
1 = inclusion criteria
2 = no mention

Starting Date
8. Did the authors select a year as a starting date?

Variable Name: yr_start
Variable Type: Nominal

Code:
0 = no
1 = yes

(If the answer is NO, skip to question 9)

8a. If the authors selected a year as the starting date, what reason was given for choosing this date?

Variable Name: yr_start_why
Variable Type: Nominal

Code:
1 = first year of test publication (stated)
2 = year new test form was published
3 = set to make the RG study manageable
4 = none given
5 = first year of test publication (not explicitly stated)
6 = other

(if the answer is other, proceed to 8aa otherwise skip to question 9)

8aa. If other reasons were given, specify

Variable name: other_reason
Variable type: String
RG CHOICES: Coding the Test and Sample:

Coding Test Characteristics
When coding the studies for RG analysis, was this characteristic of the measurement instrument coded for?

9a. Did response scale type vary across primary studies included in RG
   Variable name: repscalevaried
   Variable type: nominal
   **Code:** 0 = no
   1 = yes
   *(If the answer is NOT VARIED, skip to 9b)*

9aa. Coding Test Characteristics: Type of response scale
    Variable Name: respscaletyp
    Variable Type: Nominal
    **Code:** 0 = not coded
    1 = coded

9aaa. Coding Test Characteristics: How was the type of response scale coded?
    Variable Name: respscaletyphow
    Variable Type: String

9aaaa. If response scale was coded, was it analyzed
    Variable Name: repscaleanalyzed
    Variable Type: Nominal
    **Code:** 0 = not used in analysis
    1 = used in analysis

9b. Did who completed the measure vary across primary studies included in RG
   Variable name: whocompvaried
   Variable type: nominal
   **Code:** 0 = no
   1 = yes
   *(If the answer is NO, skip to 9c)*

9bb. Coding Test Characteristics: Who completed the measure?
    Variable Name: whocomp
    Variable Type: Nominal
    **Code:** 0 = not coded
    1 = coded

9bbb. Coding Test Characteristics: How did they code who completed the measure?
    Variable Name: whocomphow
    Variable Type: String
9b. If ‘who completed the measure’ was coded, was it analyzed
   Variable Name: whocompanalyzed
   Variable Type: Nominal
   Code:   0 = not used in analysis
           1 = used in analysis

9c. Were multiple translations of the instrument included in the RG study?
   Variable name: languagevaried
   Variable type: nominal
   Code:   0 = no
           1 = yes
   (If the answer is NO, skip to 9d)

9cc. Coding Test Characteristics: language of test (i.e. between English and Japanese)
   Variable Name: langtest
   Variable Type: Nominal
   Code:   0 = not coded
           1 = coded

9ccc. If multiple translations were coded, was it analyzed
   Variable Name: languageanalyzed
   Variable Type: Nominal
   Code:   0 = not used in analysis
           1 = used in analysis

9d. Coding Test Characteristics: Form of test used (i.e. Original form, revised form, adapted form)
   Variable Name: original
   Variable Type: Nominal
   Code:   0 = not coded
           1 = coded
   (If the answer is NO, skip to 10a)

9ddd. If multiple forms were coded, was it analyzed
   Variable Name: formsanalyzed
   Variable Type: Nominal
   Code:   0 = not used in analysis
           1 = used in analysis
Coding sample characteristics:
Each of the following are characteristics of the sample that RG researchers may choose to use in RG analysis.

Were the following coded for in this particular RG study?

10a. Coding Sample Characteristics: Sample Size
   Variable Name: sampsize
   Variable Type: Nominal
   Code: 0 = no
        1 = yes
   (If the answer is NO, skip to 10b)

10aa. If sample size was coded, was it analyzed
   Variable Name: sampsizeanalyzed
   Variable Type: Nominal
   Code: 0 = not used in analysis
        1 = used in analysis

10aaa. Notes on sample size
   Variable name: sampsizenotes
   Variable type: String

10b. Coding Sample Characteristics: gender homogeneity/proportion
   Variable Name: sex
   Variable Type: Nominal
   Code: 0 = no
        1 = yes
   (If the answer was NO, skip to 10c)

10bb. Coding Sample Characteristics: If gender homogeneity was coded, how?
   Variable Name: sex_how
   Variable Type: String

10bbb. If gender was coded, was it analyzed
   Variable Name: genderanalyzed
   Variable Type: Nominal
   Code: 0 = not used in analysis
        1 = used in analysis

10c. Coding Sample Characteristics: racial/ethnic homogeneity/proportion
   Variable Name: race
   Variable Type: Nominal
   Code: 0 = no
        1 = yes
   (If the answer was NO, skip to 10d)
10cc. Coding Sample Characteristics: If race/ethnicity is coded, how was it coded?
   Variable Name: race_how
   Variable Type: String

10ccc. If race/ethnicity was coded, was it analyzed
   Variable Name: raceanalyzed
   Variable Type: Nominal
   Code: 0 = not used in analysis
         1 = used in analysis

10cccc. Notes on race/ethnicity
   Variable name: racenotes
   Variable type: String

10d. Coding Sample Characteristics: coded for age?
   Variable Name: age
   Variable Type: Nominal
   Code: 0 = no
         1 = yes
   (If the answer is NO, skip to 10e)

10dd. Coding Sample Characteristics: If age was coded for, how was this done?
   Variable Name: age_how
   Variable Type: Nominal
   Code: 0 = continuous variable
         1 = group variable
         2 = mean age

10ddd. Coding Sample Characteristics: If age was coded by group, how were they grouped?
   Variable Name: age_grpd_how
   Variable Type: String

10dddd. If age was coded, was it analyzed
   Variable Name: genderanalyzed
   Variable Type: Nominal
   Code: 0 = not used in analysis
         1 = used in analysis

10e. Coding Sample Characteristics: population type (i.e. clinical or not clinical)
   Variable Name: poptype
   Variable Type: Nominal
   Code: 0 = no
         1 = yes
   (If the answer is NO, skip to 10f)
10ee. Coding Sample Characteristics: If population type was coded, how was it coded?
   Variable Name: poptype_how
   Variable Type: String

10eee. If population type was coded, was it analyzed
   Variable Name: genderanalyzed
   Variable Type: Nominal
   Code: 0 = not used in analysis
       1 = used in analysis

10eee. Notes on population
   Variable name: popnotes
   Variable type: String

10f. Coding Sample Characteristics: What other sample characteristics were coded for?
   Variable Name: other_char
   Variable Type: String

10ff. If other characteristics were coded, were they analyzed (note each one)
   Code: 0 = not used in analysis
       1 = used in analysis

10g. Coding Sample Characteristics: List any characteristics the authors wanted to
    analyze but could not due to lack of reporting.
   Variable Name: wantedanalyze
   Variable Type: String

Interrater Reliability:

RG researchers Interrater Reliability for coding
11. IRR for Coding: Did the authors do Inter-Rater Reliability for coding?
    Variable Name: irr_coded
    Variable Type: Nominal
    Code: 0 = no
        1 = yes
        (If the answer to question 11 is NO, skip to question 13a)

11a. IRR for Coding: If authors did check IRR, how many raters were used?
    Variable Name: num_irr
    Variable Type: Numeric

Methods of Computing Inter-rater Reliability
The following, questions 12a through 12d, all deal with how inter-rater reliability
was computed. If IRR was not computed, skip to question 13.
Independent and Dependent Variables:

Reliability used in Analysis
This question addresses what type(s) of reliability was(were) used as an outcome in the RG analysis. Although multiple types of reliability may have been reported in the primary articles used by the RG author(s), only those types of reliability which the RG author(s) used in the ACTUAL ANALYSIS should be coded here.

12a. Method of computing IRR: percent agree
Variable Name: per_agree
Variable Type: Nominal
Code: 0 = no
1 = yes

12b. Method of computing IRR: Cohen's Kappa
Variable Name: kappa
Variable Type: Nominal
Code: 0 = no
1 = yes

12c. Method of computing IRR: Intra-class correlation
Variable Name: icc
Variable Type: Nominal
Code: 0 = no
1 = yes

12d. Method of computing IRR: Not given
Variable Name: meth_unknown
Variable Type: Nominal
Code: 0 = no
1 = yes

13a. Reliability Examined: IRR (inter-rater reliability)
Variable Name: irr
Variable Type: Nominal
Code: 0 = not used
1 = used

13b. Reliability Examined: KR-20
Variable Name: kr20
Variable Type: Nominal
Code: 0 = not used
1 = used
13c. Reliability Examined: Alpha
   Variable Name: alpha
   Variable Type: Nominal
   Code: 0 = not used
         1 = used

13d. Reliability Examined: Split-half
   Variable Name: splithalf
   Variable Type: Nominal
   Code: 0 = not used
         1 = used

13e. Reliability Examined: Test-retest
   Variable Name: testretest
   Variable Type: Nominal
   Code: 0 = not used
         1 = used

13f. Reliability Examined: Parallel forms
   Variable Name: parallel
   Variable Type: Nominal
   Code: 0 = not used
         1 = used

13g. If applicable, list any types of reliability that were dropped from the more sophisticated analyses – and why
   Variable Name: reliabilitydropped
   Variable Type: String

Coding Factors that affect Reliability:
The RG author often codes for aspects of the test instrument and primary analyses that may affect reliability of the scores when the instrument is used.

14a. Were measures (or versions) of varying lengths included?
    Variable Name: lengthvaried
    Variable type: nominal
    Code: 0 = no
          1 = yes
    (If the answer to question 14a is NO, skip to question 15a)

14aa. Factors affecting Reliability: If measure/length varied, was length coded for?
      Variable Name: testing
      Variable Type: Nominal
      Code: 0 = not coded for
             1 = coded for
14aaa. If test length was coded, was it analyzed
   *Variable Name: testanalyzed*
   *Variable Type: Nominal*
   *Code: 0 = not used in analysis*
   *1 = used in analysis*

14b. Factors affecting Reliability: Correction for attenuation
   *Variable Name: coratten*
   *Variable Type: Nominal*
   *Code: 0 = not coded for*
   *1 = coded for*

14c. Factors affecting Reliability: Standard Deviation of Scale
   *Variable Name: sdscale*
   *Variable Type: Nominal*
   *Code: 0 = not coded for*
   *1 = coded for*

14cc. Did they use SD or variance as a predictor?
   *Variable Name: SD_variance_analyzed*
   *Variable Type: Nominal*
   *Code: 0 = not used in analysis*
   *1 = used in analysis*

14ccc. Did they use a weighting approach?
   *Variable Name: weight*
   *Variable Type: Nominal*
   *Code: 0 = not used*
   *1 = used*

   *(If the answer to question 14ccc is NOT USED, skip to question 14d)*

14cccc. What type of weighting approach did they use (specify N, SD, both, or other)?
   *Variable name: typeweight*
   *Variable type: String*

14d. Factors affecting Reliability: Did the authors report the reliability of:
   *Variable Name: subgroup*
   *Variable Type: Nominal*
   *Code: 1 = only subgroups (i.e. gender)*
   *2 = only whole groups*
   *3 = both subgroups (i.e. gender) and whole groups*
Addressing Recognized Problems: Independence

Reliability Independence Issues: Reliability can be non-independent when multiple subscales are used with the same group, multiple subgroups are used from one study, multiple measures are used, and/or multiple reliability coefficients are used from the same study. The following set of questions begins with a branching question: if all reliability reports are completely independent, skip to question 16 after answering question 15.

15. Reliability Independence Issues: were all the reliability reports independent
   
   **Variable Name:** independent
   **Variable Type:** Nominal
   **Code:**
   - 0 = no
   - 1 = yes
   
   *(If answer is yes, skip to question 16)*

   Reasons for non-independence:
   15a. Independence Issues: Reported reliability for more than 1 subscale
       
       **Variable Name:** ind_subscale
       **Variable Type:** Nominal
       **Code:**
       - 0 = no
       - 1 = yes
       
       *(If the answer is NO, skip to question 15b)*

   15aa. Independence Issues: If multiple subscale reliabilities WERE reported in primary studies, how did RG author(s) handle this during data analysis?
       
       **Variable Name:** ind_subscale__method
       **Variable Type:** Nominal
       **Code:**
       - 1 = conducted separate analysis by subscale (specify)
       - 2 = combined multiple subscale reliabilities in one analysis
       - 3 = only used one type of subscale
       - 4 = used a fixed effect MANOVA
       - 5 = used HLM multivariate statistics
       - 6 = dummy coded
       - 7 = did not address the issue in any way (verbal or data analysis)
       - 8 = issue verbally addressed but did not address it in the analysis
       - 9 = other (specify)

   15aaa. If 15aa response was 1, specify
       
       **Variable name:** separatesubscale
       **Variable type:** string

   15aaaa. If 15aa response was 2, did they justify it (8) or not address the issue (7)
       
       **Variable name:** combinedsubscale
       **Variable type:** Nominal
       **Code:**
       - 7 = did not address the issue in any way (verbal or data analysis)
       - 8 = issue verbally addressed but did not address it in the analysis
15aaaaa. If 15aaaa response was 8 – what is the reason
   Variable name: combinedsubscalejustification
   Variable type: string

15b. Independence Issues: More than one subgroup reliability reported per study
   Variable Name: ind_subgrp
   Variable Type: Nominal
   Code: 0 = no
   1 = yes
   (If the answer is NO, skip to question 15c)

15bb. Independence Issues: If multiple subgroup reliabilities WERE reported in primary
       studies, how did RG author(s) handle this during data analysis?
       Variable Name: ind_subgrp_method
       Variable Type: Nominal
       Code: 1 = conducted separate analysis by subgroup
             2 = combined multiple subgroups in one analysis
             3 = only used one subgroup
             4 = used a fixed effect MANOVA
             5 = used HLM multivariate statistics
             6 = dummy coded
             7 = did not address the issue in any way (verbal or data analysis)
             8 = issue verbally addressed but did not address it in the analysis
             9 = other (specify)

15bbb. If 15bb response was 2, did they justify it (8) or not address the issue (7)
       Variable name: combinedsubgroup
       Variable type: Nominal
       Code: 7 = did not address the issue in any way (verbal or data analysis)
             8 = issue verbally addressed but did not address it in the analysis

15bbbb. If 15bbbb response was 8 – what is the reason
       Variable name: combinedsubgroupjustification
       Variable type: string

15c. Independence Issues: Used multiple measures (i.e. including SAT and ACT in 1
     study)
     Variable Name: ind_measures
     Variable Type: Nominal
     Code: 0 = no
     1 = yes
     (If the answer is NO, skip to question 15d)
15cc. Independence Issues: If multiple measures WERE reported in primary studies, how did RG author(s) handle this during data analysis?

Variable Name: ind_measures_method
Variable Type: Nominal

Code: 1 = conducted separate analysis by measure (specify)
2 = combined multiple measures in one analysis
3 = only used one measure
4 = used a fixed effect MANOVA
5 = used HLM multivariate statistics
6 = dummy coded
7 = did not address the issue in any way (verbal or data analysis)
8 = issue verbally addressed but did not address it in the analysis
9 = other (specify)

15ccc. If 15cc response was 1, specify
Variable name: separatemeasure
Variable type: string

15cccc. If 15cc response was 2, did they justify it (8) or not address the issue (7)

Variable name: combinedmeasure
Variable type: Nominal

Code: 7 = did not address the issue in any way (verbal or data analysis)
8 = issue verbally addressed but did not address it in the analysis

15ccccc. If 15cccc response was 8 – what is the reason
Variable name: combinedmeasureustification
Variable type: string

15d. Independence Issues: More than one type of reliability (i.e. alpha and test-retest) reported per study

Variable Name: ind_rel
Variable Type: Nominal

Code: 0 = no
1 = yes

(If the answer is NO, skip to question 16)
15dd. Independence Issues: If multiple types of reliability WERE reported in primary studies, how did RG author(s) handle this during data analysis?

Variable Name: ind_rel_method
Variable Type: Nominal
Code: 1 = conducted separate analysis by reliability type (specify)
  2 = combined multiple types of reliability in one analysis
  3 = only used one type of reliability estimate
  4 = used a fixed effect MANOVA
  5 = used HLM multivariate statistics
  6 = dummy coded
  7 = did not address the issue in any way (verbal or data analysis)
  8 = issue verbally addressed but did not address it in the analysis
  9 = other (specify)
 10 = 1 and 2

15ddd. If 15dd response was 1 or 10, specify

Variable name: separatetype
Variable type: string

15dddd. If 15aa response was 2 or 10, did they justify it (8) or not address the issue (7)

Variable name: combinedtype
Variable type: Nominal
Code: 7 = did not address the issue in any way (verbal or data analysis)
  8 = issue verbally addressed but did not address it in the analysis

15ddddd. If 15aaaa response was 8 – what is the reason

Variable name: combinedtypejustification
Variable type: Nominal

Addressing Recognized Problems: Missing Data

Other Factors affecting Reliability:
16. Factors affecting Reliability: Did they check for Publication bias or file drawer bias?

Variable Name: pubbias
Variable Type: Nominal
Code: 0 = no
  1 = yes
(If answer is NO, skip to question 17)

16a. Factors affecting Reliability: If publication bias was checked, what method was used?

Variable Name: pubbias_method
Variable Type: Nominal
Code: 0 = method not given
  1 = Begg’s test
  2 = funnel-plot technique and failsafe n
  3 = other (specify)
16aa. Additional details of how publication bias was checked
  Variable name: pubbias_details
  Variable type: string

Dealing with Missing Data:
Some primary studies do not include reliability information, which is the dependent variable in most RG studies. RG researchers can ignore this data or attempt to acknowledge or correct for the missing reliability information. The following questions deal with this issue. If the answer to 17 is NO, skip to question 18.

17. Missing Data: Did the authors include studies that did not report reliability?
  Variable Name: missing
  Variable Type: Nominal
  Code: 0 = no
  1 = yes
  (if NO, skip to question 18)

17a. Did the measure(s) use a dichotomous scale?
  Variable Name: dichotomous
  Variable Type: Nominal
  Code: 0 = no
  1 = yes

17b. Missing Data: Did the authors use KR-21?
  Variable Name: kr21
  Variable Type: Nominal
  Code: 0 = no
  1 = yes

17c. Missing Data: Did authors e-mail article authors for reliability information?
  Variable Name: email
  Variable Type: Nominal
  Code: 0 = no
  1 = yes

17d. Missing Data: If other methods were used to include studies with missing data, what was done?
  Variable Name: miss_other
  Variable Type: String
Transformations, Data Used for Analysis, Power, and Homogeneity of Variances

Transformations
18. Transformations: Did authors transform the data for analysis?
   Variable Name: transform
   Variable Type: Nominal
   Code: 0 = no
   1 = yes
   (If the answer is NO, skip to question 19)

Data used for analysis (more than one can apply)
18.5a. Did the authors use raw data for the reliability coefficients
   Variable Name: rawdata
   Variable Type: Nominal
   Code: 0 = not used
   1 = used

18.5b. Did the authors use the square root of reliability for the reliability coefficients
   Variable Name: squareroot
   Variable Type: Nominal
   Code: 0 = not used
   1 = used

18.5c. Did the authors use fisher’s r-to-z for the reliability coefficients
   Variable Name: fisher
   Variable Type: Nominal
   Code: 0 = not used
   1 = used

18.5d. Did the authors use the squared reliability coefficient for the reliability coefficients
   Variable Name: squared
   Variable Type: Nominal
   Code: 0 = not used
   1 = used

18.5e. If a different method was used specify
   Variable Name: data_used_other
   Variable Type: String

Power
19. Power: Did authors conduct an a priori power analysis?
   Variable Name: power
   Variable Type: Nominal
   Code: 0 = no
   1 = yes
Homogeneity of Variances
Question 20 is a branching question. If the answer is NO, skip to question 21 after answering 20.

20. Homogeneity: Did the authors conduct a test for homogeneity of population correlations?
   Variable Name: homogtest
   Variable Type: Nominal
   Code: 0 = no
   1 = yes
   (If the answer is NO, skip to question 21)

20a. Homogeneity: Homogeneity was tested with an HLM random effects test
   Variable Name: hlm
   Variable Type: Nominal
   Code: 0 = no
   1 = yes

20b. Homogeneity: Homogeneity was tested with a Q test
   Variable Name: qtest
   Variable Type: Nominal
   Code: 0 = no
   1 = yes

20c. Homogeneity: Homogeneity was tested with a different method
   Variable Name: homog_other
   Variable Type: Nominal
   Code: 0 = no
   1 = yes
   (If answer is NO, skip to question 21)

20cc. Homogeneity: If homogeneity was tested using a different method, describe.
   Variable Name: homog_othersp
   Variable Type: String

Data Analysis, CI, Moderators, and Variance

Data Analysis
21. Data analysis: Type of data analysis
   Variable Name: DataAnalysis
   Variable Type: Nominal
   Code: 0 = fixed effects
   1 = random effects
   2 = HLM
   3 = descriptive (Primary analysis)
   4 = correlations (primary analysis)
   5 = other (specify)
21a. If other in 21, specify
   Variable name: otherdatanalysis
   Variable type: String

22. Data analysis: Did researchers justify the type of data analysis used?
   Variable Name: justifyDA
   Variable Type: Nominal
   Code: 0 = no
       1 = yes
   (If the answer is YES, skip to question 23)

22a. Data analysis: If authors justified the data analysis, what was the justification?
   Variable Name: justifyDAgiven
   Variable Type: String

Confidence Intervals
23. Confidence Intervals: Did the authors report CI?
   Variable Name: reportCI
   Variable Type: Nominal
   Code: 0 = no
       1 = yes
   (If the answer is NO, skip to question 24)

23a. Confidence Intervals: If CIs were reported, what were the CIs based upon?
   Variable Name: typeCI
   Variable Type: Nominal
   Code: 1 = fixed model
         2 = random/mixed effects models
         Missing = No CIs were reported

Numbers
24. Number of RG articles found by authors with reported RG estimates for the primary studies
   Variable Name: numart
   Variable type: Numeric

24. Notes on primary studies
   Variable name: studynotes
   Variable type: string

25. Total number of observed reliability estimates used in the RG data analysis
   Variable Name: Numrel
   Variable Type: Numeric
26. Number of reliability estimates for Measure 1  Variable Name: IC1  
   Variable Name: OverallIC  
   Variable Type: Numeric

26.25 If measure 1 IC had multiple types of IC – how many were in each type  
   Variable Name: ICbreakdown  
   Variable Type: Numeric

26.50. Notes on the Measure 1 IC reliability estimates  
   Variable Name: ICnotes  
   Variable Type: String

26a. Number of reliability estimates for Measure 1 – Subscale 1– Internal Consistency  
   Variable Name: IC1.1  
   Variable Type: Numeric

26b. Number of reliability estimates for Measure 1 – Subscale 2– Internal Consistency  
   Variable Name: IC1.2  
   Variable Type: Numeric

26c. Number of reliability estimates for Measure 1 – Subscale 3– Internal Consistency  
   Variable Name: IC1.3  
   Variable Type: Numeric

26d. Number of reliability estimates for Measure 1 – Subscale 4– Internal Consistency  
   Variable Name: IC1.4  
   Variable Type: Numeric

26e. Number of reliability estimates for Measure 1 – Subscale 5– Internal Consistency  
   Variable Name: IC1.5  
   Variable Type: Numeric

26f. Number of reliability estimates for Measure 1 – Subscale 6– Internal Consistency  
   Variable Name: IC1.6  
   Variable Type: Numeric

26g. Number of reliability estimates for Measure 1 – Subscale 7– Internal Consistency  
   Variable Name: IC1.7  
   Variable Type: Numeric

26h. Number of reliability estimates for Measure 1 – Subscale 8– Internal Consistency  
   Variable Name: IC1.8  
   Variable Type: Numeric
26i. Number of reliability estimates for Measure 1 – Subscale 9– Internal Consistency
   Variable Name: IC1.9
   Variable Type: Numeric

26j. Number of reliability estimates for Measure 1 – Subscale 10– Internal Consistency
   Variable Name: IC1.10
   Variable Type: Numeric

26k. Number of reliability estimates for Measure 1 – Subscale 11– Internal Consistency
   Variable Name: IC1.11
   Variable Type: Numeric

27. Number of reliability estimates for Measure 1 – Test-Retest
   Variable Name: TR_T1
   Variable Type: Numeric

27.50. Notes on the Measure 1 Test-retest reliability estimates
   Variable Name: T-RTnotes
   Variable Type: String

27a. Number of reliability estimates for Measure 1 – Subscale 1 - Test-Retest
   Variable Name: TR_T1.1
   Variable Type: Numeric

27b. Number of reliability estimates for Measure 1 – Subscale 2 - Test-Retest
   Variable Name: TR_T1.2
   Variable Type: Numeric

27c. Number of reliability estimates for Measure 1 – Subscale 3 - Test-Retest
   Variable Name: TR_T1.3
   Variable Type: Numeric

27d. Number of reliability estimates for Measure 1 – Subscale 4 - Test-Retest
   Variable Name: TR_T1.4
   Variable Type: Numeric

27e. Number of reliability estimates for Measure 1 – Subscale 5 - Test-Retest
   Variable Name: TR_T1.5
   Variable Type: Numeric

27f. Number of reliability estimates for Measure 1 – Subscale 6 - Test-Retest
   Variable Name: TR_T1.5
   Variable Type: Numeric

27g. Number of reliability estimates for Measure 1 – Subscale 7 - Test-Retest
   Variable Name: TR_T1.5
   Variable Type: Numeric
27h. Number of reliability estimates for Measure 1 – Subscale 8 - Test-Retest  
*Variable Name: TR_T1.5*  
*Variable Type: Numeric*

27i. Number of reliability estimates for Measure 1 – Subscale 9 - Test-Retest  
*Variable Name: TR_T1.5*  
*Variable Type: Numeric*

27j. Number of reliability estimates for Measure 1 – Subscale 10 - Test-Retest  
*Variable Name: TR_T1.5*  
*Variable Type: Numeric*

28. Number of reliability estimates for Measure 2 – Internal Consistency (alpha & KR-20)  
*Variable Name: IC2*  
*Variable Type: Numeric*

28.5. Notes on the Measure 2 IC reliability estimates  
*Variable Name: IC2notes*  
*Variable Type: String*

28a. Number of reliability estimates for Measure 2 – Subscale 1– Internal Consistency  
*Variable Name: IC2.1*  
*Variable Type: Numeric*

28b. Number of reliability estimates for Measure 2 – Subscale 2– Internal Consistency  
*Variable Name: IC2.2*  
*Variable Type: Numeric*

28c. Number of reliability estimates for Measure 2 – Subscale 3– Internal Consistency  
*Variable Name: IC2.3*  
*Variable Type: Numeric*

28d. Number of reliability estimates for Measure 2 – Subscale 4– Internal Consistency  
*Variable Name: IC2.4*  
*Variable Type: Numeric*

28e. Number of reliability estimates for Measure 2 – Subscale 5– Internal Consistency  
*Variable Name: IC2.5*  
*Variable Type: Numeric*

28f. Number of reliability estimates for Measure2 – Subscale 6– Internal Consistency  
*Variable Name: IC2.6*  
*Variable Type: Numeric*
29. Number of reliability estimates for Measure 2 – Test-Retest
   \textit{Variable Name: TR\textsubscript{T2}}
   \textit{Variable Type: Numeric}

29a. Number of reliability estimates for Measure 2 – Subscale 1 - Test-Retest
   \textit{Variable Name: TR\textsubscript{T2.1}}
   \textit{Variable Type: Numeric}

29b. Number of reliability estimates for Measure 2 – Subscale 2 - Test-Retest
   \textit{Variable Name: TR\textsubscript{T2.2}}
   \textit{Variable Type: Numeric}

29c. Number of reliability estimates for Measure 2 – Subscale 3 - Test-Retest
   \textit{Variable Name: TR\textsubscript{T2.3}}
   \textit{Variable Type: Numeric}

29d. Number of reliability estimates for Measure 2 – Subscale 4 - Test-Retest
   \textit{Variable Name: TR\textsubscript{T2.4}}
   \textit{Variable Type: Numeric}

30. Number of reliability estimates for Measure 3 – Internal Consistency (alpha & KR-20)
   \textit{Variable Name: IC3}
   \textit{Variable Type: Numeric}

30.50. Notes on the Measure 4 IC reliability estimates
   \textit{Variable Name: IC3notes}
   \textit{Variable Type: String}

30a. Number of reliability estimates for Measure 3 – Subscale 1– Internal Consistency
   \textit{Variable Name: IC3.1}
   \textit{Variable Type: Numeric}

30b. Number of reliability estimates for Measure 3 – Subscale 2– Internal Consistency
   \textit{Variable Name: IC3.2}
   \textit{Variable Type: Numeric}

30c. Number of reliability estimates for Measure 3 – Subscale 3– Internal Consistency
   \textit{Variable Name: IC3.3}
   \textit{Variable Type: Numeric}

30d. Number of reliability estimates for Measure 3 – Subscale 4– Internal Consistency
   \textit{Variable Name: IC3.4}
   \textit{Variable Type: Numeric}
30e. Number of reliability estimates for Measure 3 – Subscale 5– Internal Consistency
   Variable Name: IC3.5
   Variable Type: Numeric

30f. Number of reliability estimates for Measure 3 – Subscale 6– Internal Consistency
   Variable Name: IC3.6
   Variable Type: Numeric

31. Number of reliability estimates for Measure 3 – Test-Retest
   Variable Name: TR_T3
   Variable Type: Numeric

31a. Number of reliability estimates for Measure 3 – Subscale 1 - Test-Retest
   Variable Name: TR_T3.1
   Variable Type: Numeric

31b. Number of reliability estimates for Measure 3 – Subscale 2 - Test-Retest
   Variable Name: TR_T3.2
   Variable Type: Numeric

31c. Number of reliability estimates for Measure 3 – Subscale 3 - Test-Retest
   Variable Name: TR_T3.3
   Variable Type: Numeric

31d. Number of reliability estimates for Measure 3 – Subscale 4 - Test-Retest
   Variable Name: TR_T3.4
   Variable Type: Numeric

32. Number of reliability estimates for Measure 4 – Internal Consistency (alpha & KR-20)
   Variable Name: IC4
   Variable Type: Numeric

32a. Number of reliability estimates for Measure 4 – Subscale 1– Internal Consistency
   Variable Name: IC4.1
   Variable Type: Numeric

32b. Number of reliability estimates for Measure 4 – Subscale 2– Internal Consistency
   Variable Name: IC4.2
   Variable Type: Numeric

32c. Number of reliability estimates for Measure 4 – Subscale 3– Internal Consistency
   Variable Name: IC4.3
   Variable Type: Numeric
32d. Number of reliability estimates for Measure 4 – Subscale 4 – Internal Consistency  
   Variable Name: IC4.4  
   Variable Type: Numeric

33. Number of reliability estimates for Measure 4 – Test-Retest  
   Variable Name: TR_T4  
   Variable Type: Numeric

33a. Number of reliability estimates for Measure 4 – Subscale 1 - Test-Retest  
   Variable Name: TR_T4.1  
   Variable Type: Numeric

33b. Number of reliability estimates for Measure 4 – Subscale 2 - Test-Retest  
   Variable Name: TR_T4.2  
   Variable Type: Numeric

33c. Number of reliability estimates for Measure 4 – Subscale 3 - Test-Retest  
   Variable Name: TR_T4.3  
   Variable Type: Numeric

33d. Number of reliability estimates for Measure 4 – Subscale 4 - Test-Retest  
   Variable Name: TR_T4.4  
   Variable Type: Numeric

34. Number of reliability estimates for Measure 5 – Internal Consistency (alpha & KR-20)  
   Variable Name: IC5  
   Variable Type: Numeric

34a. Number of reliability estimates for Measure 5 – Subscale 1 – Internal Consistency  
   Variable Name: IC5.1  
   Variable Type: Numeric

34b. Number of reliability estimates for Measure 5 – Subscale 2 – Internal Consistency  
   Variable Name: IC5.2  
   Variable Type: Numeric

34c. Number of reliability estimates for Measure 5 – Subscale 3 – Internal Consistency  
   Variable Name: IC5.3  
   Variable Type: Numeric

34d. Number of reliability estimates for Measure 5 – Subscale 4 – Internal Consistency  
   Variable Name: IC5.4  
   Variable Type: Numeric
35. Number of reliability estimates for Measure 5 – Test-Retest
   
   Variable Name: TR_T5
   Variable Type: Numeric

35a. Number of reliability estimates for Measure 5 – Subscale 1 - Test-Retest
   
   Variable Name: TR_T5.1
   Variable Type: Numeric

35b. Number of reliability estimates for Measure 5 – Subscale 2 - Test-Retest
   
   Variable Name: TR_T5.2
   Variable Type: Numeric

35c. Number of reliability estimates for Measure 5 – Subscale 3 - Test-Retest
   
   Variable Name: TR_T5.3
   Variable Type: Numeric

35d. Number of reliability estimates for Measure 5 – Subscale 4 - Test-Retest
   
   Variable Name: TR_T5.4
   Variable Type: Numeric

32. Number of reliability estimates for Measure 5 – Interrater reliability
   
   Variable Name: IRR5
   Variable Type: Numeric

* Note: One study had a total of 22 measures; therefore, the coding spreadsheet includes variables 35.01-35.17 to have a column for each measure, but for space reasons each of these variables is not included here

**Moderators**

36. Moderators: Did authors test for moderators in DA
   
   Variable Name: moderator
   Variable Type: Nominal
   Code: 0 = no
   1 = yes
   (If the answer is NO, skip to question 25)

36a. Moderators were tested for in what way?
   
   Variable Name: mod_how
   Variable Type: Nominal
   Code: 1 = test moderators all at once
   2 = test moderators separately
   3 = test moderators in a series of blocks or hierarchical regression methods
   4 = other (specify)
   5 = tested, but not reported how
36aa. Moderator notes
   Variable name: moderatornotes
   Variable type: string

Variability in Reliability Coefficients
37a. Variability in reliability coefficients: Box and Whisker plot used?
   Variable Name: boxwhisker
   Variable Type: Nominal
   Code: 0 = no
         1 = yes

37b. Variability in reliability coefficients: Means and SD used?
   Variable Name: meansd
   Variable Type: Nominal
   Code: 0 = no
         1 = yes

Reporting reliability in primary studies as reported by RG researchers. If the number is given, report the number. If the number is not given, report the percent

38. Did the RG researchers give the reports of reliability from the primary studies
   Variable name: relreport
   Variable type: nominal
   Code: 0 = no
         1 = yes
   (If the answer to 38 is NO, skip the rest of the questions given below)

38a. Number of primary studies that did not mention reliability or did not report reliability
   Variable name: nomention
   Variable type: numeric

38b. Number of primary studies that mentioned reliability but did not cite any values
   Variable name: justmention
   Variable type: numeric

38c. Number of primary studies that did reliability induction
   Variable name: induction
   Variable type: numeric

38d. Number of primary studies that reported reliability for their sample
   Variable name: samplerel
   Variable type: numeric
38e. If another category of reports of the reliability from the primary studies was given, please describe the category and give the number in the category

*Variable name: otherreport
*Variable type: string*

38f. Notes on reporting reliability

*Variable name: reliabilityreportnotes
*Variable type: string*

*Note: If a variable does not provide information, or information was not given when it should have been provided, the code 888 is used; for example if the RG researchers said they coded for age but did not specify how age was coded
*Note: If something is not applicable it is coded as 999
References


generalization: A strategy to compute a fail-safe N with reliability coefficients.
*Educational and Psychological Measurement, 68*, 120-128.
doi:10.1177/0013164407301528


Thompson, B. (1999, January). *Reliability generalization: An important meta-analysis method, because it is incorrect to say “the test is unreliable.”* Paper presented at the meeting of the Southwest Educational Research Association, San Antonio, TX.


Vita

Author’s Name: Alexandra Marie Henchy

Education

2007 University of Miami
Bachelor of Arts
Majors: Psychology and Religious Studies
Minor: Spanish

Professional Positions
Director of Institutional Effectiveness and Assessment at Asbury Theological Seminary (March 2013-Present)

Doctoral Research Associate for Student Affairs at the University of Kentucky (August 2009 – February 2013)

Teaching Assistant for COM 591/AAS 500 (Groups, Organizations, and Identity: Intercultural Communication in Black Greek Letter Organizations) at the University of Kentucky (Spring 2012)

Intern for ACT (Summer 2010)

Instructor of EDP 202 (Human Development and Learning) at the University of Kentucky (Fall 2003- Spring 2005)

Publications


Presentations


Henchy, A. (2012, March). A quantitative researcher’s journey to complete a qualitative project. Poster presented at the meeting of the Kentucky Psychological Association, Lexington, KY.

Henchy, A. (2011, June). Student affairs assessment teams: Engaging student affairs professionals in assessment. Presentation at the meeting of the Association for
the Assessment of Learning in Higher Education, Lexington, KY.


**Honors and Awards**

Division E – Counseling and Human Development at AERA 2010