ASSESSMENT AND REPORTING OF INTERCODER RELIABILITY
IN PUBLISHED META-ANALYSES RELATED TO
PRESCHOOL THROUGH GRADE 12 EDUCATION

A dissertation presented to

the faculty of

the College of Education of Ohio University

In partial fulfillment

of the requirements for the degree

Doctor of Philosophy

Holly Raffle
August 2006
This dissertation entitled

ASSESSMENT AND REPORTING OF INTERCODER RELIABILITY
IN PUBLISHED META-ANALYSES RELATED TO
PRESCHOOL THROUGH GRADE 12 EDUCATION

by

HOLLY RAFFLE

has been approved for

the Department of Educational Studies
and the College of Education by

Gordon P. Brooks
Assistant Professor of Educational Studies

Renée A. Middleton
Dean, College of Education
ABSTRACT

RAFFLE, HOLLY, Ph.D., August 2006, Educational Studies

ASSESSMENT AND REPORTING OF INTERCODER RELIABILITY IN
PUBLISHED META-ANALYSES RELATED TO PRESCHOOL THROUGH GRADE
12 EDUCATION (215 pp.)

Director of Thesis: Gordon P. Brooks

This work focused on issues related to data collection in meta-analysis, specifically the coding process by which coders extract data from primary studies. This work used content analysis to evaluate 118 meta-analyses reported in 99 journal articles from the educational literature to establish how educational researchers assess and report intercoder reliability. The principal investigator provides detailed information regarding the assessment of intercoder reliability using an item-by-item approach and a generalizability theory framework for the instrument used in the study. The principal investigator explored six topics related to the assessment and reporting of intercoder reliability: the use of meta-analysis in education, coders and coding strategy, coder training procedures, the code book, addressing intercoder reliability, and reporting intercoder reliability. Although the literature review established that researchers perceive coder unreliability as a threat to the internal validity of a study, of the 118 meta-analyses in the study, authors reported intercoder reliability analyses less than half (37.3%) of the time. The principal investigator presents recommendations for best practice in intercoder reliability analysis and reporting based on the study methodology and research findings.
Approved:

Gordon P. Brooks
Assistant Professor of Educational Studies
ACKNOWLEDGMENTS

I wish to extend my sincere appreciation to Sarah Ganslein and Leah Strazisar for their cooperation and the time, effort, and hard work they invested in coding the meta-analyses included in this study.

I would like to express special appreciation to my graduate advisor and dissertation chair, Dr. Gordon Brooks, for his guidance, support, and patience throughout my graduate studies and the dissertation process.

I will be forever grateful to my dissertation committee, Dr. James Bruning, Dr. George Johanson, and Dr. Ralph Martin, for their leadership, expertise, and encouragement throughout the dissertation process.

I would like to acknowledge Dr. David Keck, Dr. Donald Miles, and Dr. Christopher Simpson for their leadership, expertise, patience, and enthusiasm throughout the comprehensive examination process.

I would like to express sincere appreciation to Dr. and Mrs. Clifford C. Houk for their love and personal support throughout my journey through higher education.

To Jane R. Devine, for the love and support that only a mother could provide.

To my husband, Craig C. McCarthy, your love, guidance, tolerance, and support are immeasurable. Without you, none of this would be possible.

To our son, Jake Raffle McCarthy and our unborn child, thank you for the constant reminder of what is really important in life.

Finally, nothing is possible without the grace of God (2 Corinthians 4:7).
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
</tr>
<tr>
<td>CHAPTER ONE</td>
</tr>
<tr>
<td>Introduction</td>
</tr>
<tr>
<td>Background of the Study</td>
</tr>
<tr>
<td>Statement of the Problem</td>
</tr>
<tr>
<td>Subproblems</td>
</tr>
<tr>
<td>Significance of the Study</td>
</tr>
<tr>
<td>Limitations of the Study</td>
</tr>
<tr>
<td>Delimitations</td>
</tr>
<tr>
<td>Definition of Terms</td>
</tr>
<tr>
<td>CHAPTER TWO</td>
</tr>
<tr>
<td>Review of the Literature</td>
</tr>
<tr>
<td>Measurement Issues in Meta-Analysis</td>
</tr>
<tr>
<td>Current State of Interrater Reliability Assessment</td>
</tr>
<tr>
<td>Threats to Reliability in Meta-Analysis</td>
</tr>
<tr>
<td>Intercoder Agreement versus Intercoder Reliability</td>
</tr>
<tr>
<td>Approaches to Assessing Intercoder Reliability</td>
</tr>
</tbody>
</table>
Across-the-Board Approach ................................................................. 24
Item-by-Item Approach ........................................................................ 24
Hierarchical Approach ....................................................................... 25
Assessing Intercoder Reliability: Indices of Agreement for Categorical Variables. 27
Agreement Rate ................................................................................ 28
Cohen’s kappa .................................................................................. 29
Weighted kappa ............................................................................... 34
Fleiss’ kappa .................................................................................... 37
Holsti’s C.R ....................................................................................... 39
Scott’s pi .......................................................................................... 40
Krippendorff’s alpha ........................................................................ 42
Assessing Intercoder Reliability: Correlational Indices for Quantitative Variables. 45
Pearson’s r ....................................................................................... 46
Intraclass Correlation Coefficient ...................................................... 47
Assessing Intercoder Reliability: Generalizability Theory Framework .......... 49
Generalizability Theory for Relative Decisions vs. Absolute Decisions .......... 50
A Generalizability Coefficient for a Design 2 Study with a Norm-Referenced Approach ......................................................... 51
A Generalizability Coefficient for a Design 2 Study with a Criterion-Referenced Approach .......................................................... 55
Choosing Indices of Intercoder Reliability .......................................... 57
Determining an “Acceptable” Level of Reliability ............................... 58
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Software Available to Calculate Intercoder Reliability</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>Summary and Conclusions</td>
<td>60</td>
</tr>
<tr>
<td>CHAPTER THREE</td>
<td></td>
<td>62</td>
</tr>
<tr>
<td>Methodology</td>
<td></td>
<td>62</td>
</tr>
<tr>
<td>Research Design</td>
<td></td>
<td>62</td>
</tr>
<tr>
<td>Identification of the Population</td>
<td></td>
<td>62</td>
</tr>
<tr>
<td>Instrumentation</td>
<td></td>
<td>65</td>
</tr>
<tr>
<td>Coding Process</td>
<td></td>
<td>70</td>
</tr>
<tr>
<td>Reliability Samples</td>
<td></td>
<td>73</td>
</tr>
<tr>
<td>Interrater Reliability Analysis</td>
<td></td>
<td>75</td>
</tr>
<tr>
<td>“Acceptable” Reliability</td>
<td></td>
<td>89</td>
</tr>
<tr>
<td>Data Analysis</td>
<td></td>
<td>89</td>
</tr>
<tr>
<td>CHAPTER FOUR</td>
<td></td>
<td>90</td>
</tr>
<tr>
<td>Data Analysis</td>
<td></td>
<td>90</td>
</tr>
<tr>
<td>Overview</td>
<td></td>
<td>90</td>
</tr>
<tr>
<td>Interrater Reliability Analyses</td>
<td></td>
<td>91</td>
</tr>
<tr>
<td>Analysis of the Sub-problems</td>
<td></td>
<td>101</td>
</tr>
<tr>
<td>CHAPTER FIVE</td>
<td></td>
<td>126</td>
</tr>
<tr>
<td>Summary, Conclusions, and Recommendations</td>
<td></td>
<td>126</td>
</tr>
<tr>
<td>Summary of the Study</td>
<td></td>
<td>126</td>
</tr>
<tr>
<td>Statement of the Problem</td>
<td></td>
<td>127</td>
</tr>
<tr>
<td>Summary of the Findings</td>
<td></td>
<td>130</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1</td>
<td>Sample Data</td>
<td>28</td>
</tr>
<tr>
<td>Table 2</td>
<td>Cross-tabulation of Sample Data</td>
<td>30</td>
</tr>
<tr>
<td>Table 3</td>
<td>Marginal Values for Sample Data</td>
<td>32</td>
</tr>
<tr>
<td>Table 4</td>
<td>Expanded Cross-tabulation of Sample Data</td>
<td>36</td>
</tr>
<tr>
<td>Table 5</td>
<td>Frequencies for Both Coders</td>
<td>44</td>
</tr>
<tr>
<td>Table 6</td>
<td>Analysis of Variance for Coder Scores</td>
<td>52</td>
</tr>
<tr>
<td>Table 7</td>
<td>Interrater Reliability for Items 1a – 1d and Item 2</td>
<td>93</td>
</tr>
<tr>
<td>Table 8</td>
<td>Interrater Reliability for Items 3a – 3i</td>
<td>94</td>
</tr>
<tr>
<td>Table 9</td>
<td>Interrater Reliability for Items 4a – 4d</td>
<td>95</td>
</tr>
<tr>
<td>Table 10</td>
<td>Interrater Reliability for Items 5a – 5d and 6a – 6f</td>
<td>96</td>
</tr>
<tr>
<td>Table 11</td>
<td>Interrater Reliability for Items 7a – 7x</td>
<td>97</td>
</tr>
<tr>
<td>Table 12</td>
<td>Content Area of Published Meta-Analyses</td>
<td>103</td>
</tr>
<tr>
<td>Table 13</td>
<td>Number of Studies Included in Meta-Analyses Related to Education</td>
<td>104</td>
</tr>
<tr>
<td>Table 14</td>
<td>Specific Number of Coders</td>
<td>105</td>
</tr>
<tr>
<td>Table 15</td>
<td>Number of Studies Included in the Reliability Sample</td>
<td>107</td>
</tr>
<tr>
<td>Table 16</td>
<td>Proportion of Studies Included in the Reliability Sample</td>
<td>108</td>
</tr>
<tr>
<td>Table 17</td>
<td>Description of Coders</td>
<td>109</td>
</tr>
<tr>
<td>Table 18</td>
<td>Description of Coder Training Procedures</td>
<td>110</td>
</tr>
<tr>
<td>Table 19</td>
<td>Description of the Code Book</td>
<td>113</td>
</tr>
<tr>
<td>Table 20</td>
<td>Description of Method Used to Resolve Coding Discrepancies</td>
<td>114</td>
</tr>
</tbody>
</table>
Table 21 Specific Method Used to Calculate Intercoder Reliability .............................. 118

Table 22 Range of Reported Intercoder Reliability Values............................................ 124
CHAPTER ONE

Introduction

Background of the Study

Much of the methodological research related to meta-analysis focuses on the statistical methodology used to combine studies and analyze effect size (Cooper & Hedges, 1994; Fergusson et al., 2000). However, conducting a meta-analysis involves many other stages. In *The Handbook of Research Synthesis* (1994), editors Cooper and Hedges have described meta-analysis as a five-stage process: (a) problem formation, (b) data collection (which involves searching the literature), (c) data evaluation (coding the literature), (d) analysis and interpretation, and (e) public presentation (p. 9-12). Cooper and Hedges (1994) have written that “the problems that can arise during [the data evaluation] stage are among the most challenging faced by the research synthesist” (p. 10). To mitigate such problems, Stock (1994) has called for systematic coding, where the meta-analyst makes careful decisions concerning choosing study variables, constructing coding forms, training coders, and ensuring reliable coding. Further, Bullock and Svyantek (1985) have noted that the “results of the meta-analysis rely on the adequacy of the coding scheme” (p. 113). Therefore, this work concentrates on issues related to data collection, specifically the coding process by which data is extracted from primary studies.

In meta-analysis, researchers extract and then code data from primary studies into mutually exclusive categories (Rust & Cooil, 1994). Stock (1994) has described the role
of the coder as a researcher who locates, reduces, manipulates, and transcribes information (p. 126). Therefore, meta-analytic work, like other types of research methods that require interpretation, is highly dependent upon human judgment (Matt, 1989; Sacks et al., 1987; Stock et al., 1982; Stock, 1994). For this reason, Cooper and Hedges (1994) have acknowledged that during the data evaluation stage, it is of utmost importance for the meta-analyst to assess the “trustworthiness” of the coding process (p. 11). Several authors (Glass, McGaw & Smith, 1981; Matt, 1989; Orwin & Cordray, 1985; Stock et al., 1982) also have suggested that it is imperative to take coder reliability seriously. Perhaps this is because unreliability in the coding of primary studies threatens the internal validity of a meta-analysis (Cooper, 1998; Stock et al., 1982); reliability is a necessary precondition for validity (Light, Singer, & Willett, 1990).

Although researchers perceive coder unreliability as a threat to the internal validity of a study, meta-analysts do not uniformly report coder reliability measures. Yeaton and Wortman (1993) reported that only one of the meta-analyses published in the 1983 Evaluation Studies Review Annual included a measure of coder reliability. Further, only 29% of the 21 meta-analyses published between 1986 and 1988 in the Psychological Bulletin reported a measure of intercoder reliability (Yeaton & Wortman, 1993). Sacks et al. (1987) surveyed the reporting of coder reliability in the medical literature between 1966 and 1986 and found intercoder agreement clearly measured in only 4 of 86 meta-analyses and some “spot-checking” reported in 8 other meta-analyses (p. 452).

In light of the inconsistent reporting found in their study of meta-analyses in the medical literature, Sacks et al. (1987) called for a more uniform procedure in meta-
analysis reporting. The recent Quality of Reporting of Meta-analyses (QUOROM) statement, which was drafted to address standards for improving the quality of reporting of meta-analyses of randomized controlled trials, includes a checklist to standardize meta-analysis reporting in the medical literature (Mohler et al., 2000). Movements to improve the reporting of meta-analyses are consistent with Bryant’s (1984) ascertainment that the best way to enhance external validity in research synthesis is by encouraging independent replication. Bryant (1984) further has attested that in order for independent replication to occur, meta-analysts must consistently report the procedures used for both study selection and coding.

The literature pertaining to the criteria for the reporting of coding procedures and intercoder reliability in meta-analyses is often very general and left to the researcher for interpretation. Further, there is little consensus as to what needs to be included in a meta-analysis report. Bullock and Svyantek (1985) have proposed a series of criteria for evaluating meta-analysis research. Bullock and Svyantek’s (1985) criteria for meta-analyses meeting “acceptable quality standards” require that meta-analysts include detailed documentation of: (a) the coding scheme, including estimation procedures used for missing data; (b) the methods used to resolve problems in applying the coding scheme; and (c) a rigorous assessment of interrater reliability (p. 114). Although Oliver (1987) does not directly address meta-analysis reporting, he has ascertained that intercoder reliability “needs to be established” (p. 867). Finally, Abrami, Cohen, and d’Apollonia (1988) have written that meta-analytic procedures, including the coding of study features, must be “precisely described, objective, and repeatable” (p. 163).
More recently, Mostert (1996) has proposed a set of criteria for reporting meta-analyses and applied it to a sample of published meta-analyses in special education. With respect to intercoder reliability, Mostert (1996) has recommended that researchers report three items: (a) descriptions of the coded variables (independent variables) from primary studies, (b) reported interrater reliability for study coding, and (c) reported interrater reliability for calculation of effect sizes (p. 7). L’Hommedieu, Menges, and Brinko (1998) have recommended that researchers report two items: a description of the coding procedures and a report of how the effect size was calculated. Chambers (2004) has also described criteria for evaluating meta-analyses and suggested that meta-analyses include a description of the coding scheme and a report of interrater reliability. Interestingly, the aforementioned QUOROM statement (Mohler et al., 2000), which was drafted to address standards for improving the quality of reporting of meta-analyses of randomized controlled trials, does not address intercoder reliability.

Several meta-analysis primers also offer guidelines for reporting intercoder reliability. Cooper (1998) has recommended that researchers provide details pertaining to study coding in the methods section of their reports. Specifically, Cooper (1998) has suggested that meta-analysts should “fully describe the information about each study that was collected on the coding sheets” and include information related to intercoder reliability (p. 162). Although Wolf (1986) and Glass, McGaw, and Smith (1981) do not directly address meta-analysis reporting, they have offered guidelines for practice which recommend that meta-analysts assess reliability of the coding procedures. In The Handbook of Research Synthesis, Halvorson (1994) has recommended that meta-analysts
report the following information: (a) methods used to code primary study data, (b) number of coders, (c) measures of coder agreement, if two or more investigators independently coded the data, and (d) methods used to resolve differences between coders.

To date, there has not been a work that systematically evaluated a collection of meta-analyses from the educational literature to establish how educational researchers assess and report intercoder reliability. This work hopes to fill that void by examining the variation in the reported detail of intercoder reliability in published meta-analyses related to preschool through grade 12 education. The work of Lombard, Snyder-Duch, and Braken (2002), who surveyed the assessment and reporting of intercoder reliability in content analyses from the mass communications literature, Steiner et al. (1991), who reviewed meta-analyses in organizational behaviors and human resources management, and Mostert (1996), who analyzed the reporting of meta-analyses in special education, inspired this study.

Statement of the Problem

This study attempted to answer the overarching research question: How is intercoder reliability assessed and reported in published meta-analyses related to preschool through grade 12 education?

Subproblems

This study examined 48 sub-problems in order to provide a detailed response to the research question. The sub-problems do not represent a set of criteria for judging the
quality of intercoder reliability reporting in meta-analyses. They simply serve as a tool to explore the current state of intercoder reliability reporting among published meta-analyses related to preschool through grade 12 education. The first two sub-problems relate to a general overview of the use of meta-analysis in education:

1. What content areas in education are utilizing meta-analysis?
2. What is the average number of studies included in a meta-analysis related to education?

The following nine sub-problems relate to coders and coding strategy:

3a. Did the authors report the number of coders that coded each study?
3b. How many coders rated each study?
3c. How clear (easy to find) was the information presented regarding the number of coders?
3d. What percentage of authors differentiate between the coding of the reliability sample and the coding of the actual sample?
3e. What is the average size of the reliability sample?
3f. How clear (easy to find) was the information presented regarding the size of the reliability sample?
3g. What proportion of primary studies were included in the reliability sample?
3h. Did the authors provide a description of the coders’ levels of expertise?
3i. Who coded the studies (e.g., undergraduate research assistants, graduate research assistants, study authors, etc.)?
The following four sub-problems address coder training procedures:

4a. Did the authors report coder training procedures?
4b. How much training do coders receive?
4c. How do authors describe coder training procedures?
4d. What is the criterion for consistency (e.g., stopping rule) to determine that coders were adequately trained?

The following three sub-problems address the code book:

5a. How do authors describe the code book?
5b. Do authors report how coder discrepancies were resolved?
5c. How do authors describe the procedures in place to resolve coder discrepancies?

The following six sub-problems relate to how authors address intercoder reliability:

6a. What percentage of meta-analyses acknowledge intercoder reliability?
6b. How do authors present intercoder reliability information in the text?
6c. What is the mean number of sentences devoted to the general discussion of intercoder reliability?
6d. What is the mean number of sentences devoted to the reporting of intercoder reliability analyses?
6e. What percentage of meta-analyses include at least one citation about intercoder reliability?
6f. What is the mean number of citations related to intercoder reliability?
reliability?

The following 24 sub-problems relate to the reporting of intercoder reliability:

7a. What percentage of meta-analyses report a specific method for calculating intercoder reliability?

7b. What specific methods are being reported to calculate intercoder reliability?

7c. What format are authors using to report reliability statistics?

7d. If reliability statistics are not reported, and multiple coders were used, how do authors imply coder reliability?

7e. Do authors differentiate between effect size coding and study descriptor coding?

7f. Do authors report an overall/average intercoder reliability value for a group or groups of variables related to effect size in the study?

7g. Do authors report a range of intercoder reliability for a group or groups of variables related to effect size in the study?

7h. Do authors report specific intercoder reliability values for one or more variables related to effect size in the study?

7i. Do authors report intercoder reliability values for all variables related to effect size included in the study?

7j. How clear (easy to find) was the information presented regarding the reporting of intercoder reliability of effect size variables?

7k. Do authors report an overall/average intercoder reliability value for a
group or groups of variables related to study descriptors in the study?

7l. Do authors report a range of intercoder reliability for a group or groups of variables related to study descriptors in the study?

7m. Do authors report specific intercoder reliability values for one or more variables related to study descriptors in the study?

7n. Do authors report intercoder reliability values for all variables related to study descriptors included in the study?

7o. How clear (easy to find) was the information presented regarding the reporting of intercoder reliability of study descriptor variables?

7p. If effect size and study descriptor variables are not differentiated, do authors report an overall/average intercoder reliability value for a group or groups of variables in the study?

7q. If effect size and study descriptor variables are not differentiated, do authors report a range of intercoder reliability for a group or groups of variables in the study?

7r. If effect size and study descriptor variables are not differentiated, do authors report specific intercoder reliability values for one or more variables in the study?

7s. If effect size and study descriptor variables are not differentiated, do authors report intercoder reliability values for all variables included in the study?

7t. If effect size and study descriptor variables are not differentiated, how
clear (easy to find) was the information presented regarding the reporting of intercoder reliability?

7u. What percentage of meta-analyses include a discussion of “acceptable” reliability values?

7v. What is the range of reported reliability values?

7w. How clear (easy to find) was the information presented regarding the reporting of reliability values?

7x. What is the lowest reliability criterion deemed “acceptable” by authors?

Significance of the Study

In both quantitative and qualitative research methods, there are criteria for judging the objectivity and credibility of the data collection and measurement processes (Light, Singer, & Willett, 1990; Patton, 2002). In quantitative research, researchers examine measurement quality through the principles of reliability and validity (Light, Singer, & Willett, 1990). Grounded theory, a qualitative research method that relies on coding procedures, also utilizes the principles of reliability and validity to assess research integrity (Patton, 2002). The meta-analyst is not exempt from addressing these measurement issues. Assessment of intercoder reliability is important because a coding sheet is a research instrument and therefore is subject to the same rigorous standards as instrumentation in primary research.

This study should provide a better understanding of the assessment and reporting of intercoder reliability in published meta-analyses from the educational literature. Further, this study may be helpful in creating recommendations and/or guidelines for the
assessment and reporting of intercoder reliability in meta-analysis. Ultimately, the hope is that this study enhances rigor in meta-analysis.

Limitations of the Study

The primary limitation of this study was potential intercoder unreliability. Because this study utilized a coding scheme and coders, intercoder unreliability was a potential threat to the internal validity of the study. Stock et al. (1982) have suggested that, because most data synthesis techniques rely on human judgment, the data synthesist must “grapple with the potential problem of inadequate intercoder reliability” (p. 14). Although the code book and coding sheet for this study relate to the work of Lombard, Snyder-Duch, and Braken (2002) and Steiner, Lane, Dobbins, Schnur, and McConnell (1991) and can be used with some confidence, it was duly noted that intercoder reliability must be established anew in each data synthesis (Stock et al., 1982).

Because this study was limited to published reports, it was not exempt from availability bias (also known as publication, retrieval, or selection bias), an issue often faced by meta-analysts. Hunter and Schmidt (2004) have noted that availability bias is not only “one of the most difficult and complex” aspects of meta-analysis, it is also a problem that is not limited to meta-analysis (p. 509). When statistically combining studies and analyzing effect sizes, as in a meta-analysis, there are specific methods for detecting and correcting for availability bias (Begg, 1994; Hunter & Schmidt, 2004). However, because this study was descriptive in nature, the principal investigator simply acknowledged the availability bias that is inherent to this type of research.
Delimitations

The scope of the study sample delimits this work. Some specific delimitations include:

1. Journal articles indexed in the following databases: (a) Education Resources Information Center (ERIC), a digital library of education-related resources, sponsored by the Institute of Education Sciences of the U.S. Department of Education, (b) Education Abstracts, a database produced by the H. W. Wilson Company that covers over 500 core international periodicals, monographs, and yearbooks related to contemporary education issues; and (c) PsycINFO, a database produced by the American Psychological Association that provides abstracts and citations to the scholarly literature in the behavioral sciences and mental health. The study is limited to published journal articles because dissertations often include broader reports of methodology than journal articles, which are driven by space limitation and editorial direction.

2. Journal articles written in English.

3. Journal articles published between the years 2000 and 2004. The study is limited to articles published between 2000 and 2004 because the literature review revealed that addressing intercoder reliability is a recent phenomenon, which rises from a period where many researchers were “less than rigorous in their reliability assessment” (Neuendorf, 2002, p. 142).
4. Journal articles for which one of the keywords or search terms is “meta-analysis.”

5. Journal articles from the following educational levels: early childhood education (birth through grade 3), preschool education (birth to kindergarten), elementary/secondary education (kindergarten through grade 12), elementary education (kindergarten through grades 6, 7, or 8), primary education (kindergarten through grade 3), intermediate grades (grades 4, 5, and 6), secondary education (grades 7 through 12), middle schools (grades 4 through 9), junior high schools (grades 7, 8, and 9) and high schools (grades 9 through 12).

6. Journal articles that report a study that utilized meta-analysis methodology.

7. Journal articles that can be located in the Ohio University library or through interlibrary loan.

**Definition of Terms**

*Actual sample.* The total number of research reports coded and utilized in the final analyses. The reliability sample may or may not be a subset of the actual sample.

*Availability bias.* Availability bias, also known as publication, retrieval, or selection bias, exists based on the premise that the studies available for use in a meta-analysis are a biased sample of all existing studies. Studies may not be available because they are not published (publication bias), can not be retrieved (retrieval bias), or do not
meet the inclusion criteria (selection bias) for use in the analysis (Hunter & Schmidt, 2004).

*Code book.* The purpose of a code book in meta-analysis is two-fold, to provide detailed direction to coders and to create a historical record of the coding process. According to Stock (1994), code books should contain the following information: (a) definitions of primary constructs, (b) descriptions of the general coding categories, and (c) an overview of data entry and management processes (p. 137).

*Coder.* Stock (1994) has described the role of the coder as a researcher who locates, reduces, manipulates, and transcribes information (p. 126).

*Coder training.* The purpose of coder training is to make certain that coders are adhering to the code book and utilizing the coding sheet(s) in a uniform manner (Stock, 1994).

*Coding discrepancy.* A coding discrepancy occurs when two or more coders disagree on how to code a particular item.

*Coding sheet.* Coding sheets, also known as coding forms, are where the coder records data extracted from the primary studies.

*Intercoder reliability.* Intercoder reliability is also known as interrater reliability and intercoder agreement. Intercoder reliability is a measure of the consistency of coder’s responses to items coded using the same coding sheet (Light, Singer, & Willett, 1990). Intercoder reliability is an estimate of the amount of coder error in a meta-analysis (Orwin, 1994).
*Reliability sample.* The total number of research reports coded and utilized when calculating intercoder reliability. The reliability sample may or may not be a subset of the actual sample.
CHAPTER TWO

Review of the Literature

In a 1970 address to the American Psychological Association, Senator Walter Mondale (as cited in Hunter & Schmidt, 1990, p. 35) expressed his frustration with the research process:

“… I had hoped to find research to support or to conclusively oppose my belief that quality integrated education is the most promising approach. But I have found very little conclusive evidence. For every study, statistical or theoretical, that contains a proposed solution or recommendation, there is always another, equally well documented, challenging the assumption or conclusions of the first. No one seems to agree with anyone else’s approach. But more distressing: no one seems to know what works. As a result I must confess, I stand with my colleagues confused and often disheartened... ”

Even for the most astute readers of academic journals, research remains a “generally unorganized, decentralized, and nonstandardized process without regard to how it may fit together into a comprehensive whole” (Kavale, 2001, p. 182). Fortunately, there are methods available to synthesize research by combining and comparing primary studies. The literature review is a qualitative research method of research integration that results in a narrative report of primary studies that offers an overall conclusion (Kavale, 2001). Although literature reviews are important to research, they are unable to create generalizations (Cooper & Hedges, 1994). For this reason, researchers developed meta-analysis, a more quantitative approach to research synthesis, which uses “empirically
grounded methods based on statistical significance or nonsignificance” (Kavale, 2001, p. 178). Meta-analysis is a research method that systematically and statistically combines individual research studies in effort to “discover the consistencies and account for the variability in similar-appearing studies” (Cooper & Hedges, 1994, p.4).

Although meta-analysis has proven to be a valuable method for research synthesis for the past 25 years, it is not without critics (Kavale, 2001). As such, the meta-analyst must carefully consider the standards of “objectivity, systematicity, and rigor” in order to insure the credibility of the research (Cooper & Hedges, 1994, p. 13). The idea of credibility in meta-analysis, specifically as it relates to measurement issues, is the focus of this work.

The review of the literature includes eight sections: (a) an introduction to measurement issues in meta-analysis, (b) approaches to assessing intercoder reliability, (c) assessing intercoder reliability for categorical variables, (d) assessing intercoder reliability for quantitative variables, (e) a generalizability theory framework for assessing intercoder reliability, (f) choosing indices of intercoder reliability, (g) determining an “acceptable” level of reliability, and (h) software available to calculate intercoder reliability. A final summary concludes this chapter.

Measurement Issues in Meta-Analysis

Pedhazer and Schmelkin (1991) have written about the “lamentable state” of measurement across substantive areas and emphasize that “the improvement of the quality of measures should be one of the top priorities in sociobehavioral sciences” (p. 28). Tinsley and Weiss (2000) have echoed this sentiment and suggest that “evidence
regarding both interrater reliability and interrater agreement should be reported whenever rating scales are used” (p. 117). Clearly, the meta-analyst is not exempt from addressing measurement issues, as a coding sheet is a research instrument and is accountable to the same rigorous standards as instrumentation in primary research.

With respect to meta-analysis, discussion about measurement quality is largely reserved for the reliability and validity of the dependent variable. In fact, Hunter and Schmidt (2004) have attested that “measurement error has a special status among systematic artifacts: It is the only systematic artifact that is always present in every study in a meta-analysis” (p. 95). For this reason, researchers (Hedges & Olkin, 1985; Hunter & Schmidt, 2004) have offered very specific methods for the meta-analyst to correct for the effects of measurement error. However, as described in Chapter One, several authors (Cooper, 1998; Cooper & Hedges, 1994; Glass, McGaw & Smith, 1981; Matt, 1989; Oliver, 1987; Orwin & Cordray, 1985; Stock et al., 1982; Tinsley & Weiss, 1975, 2000) have maintained that it is also important for the meta-analyst to perceive coder unreliability as a serious issue. Further, several authors (Bryant, 1984; Orwin, 1994; Sacks et al., 1987; Tinsley & Weiss, 1975, 2000; Yeaton & Wortman, 1993) have called for meta-analysts to report coder reliability procedures and measures.

Current State of Interrater Reliability Assessment

Although there is a call in the literature (Cooper, 1998; Cooper & Hedges, 1994; Glass, McGaw & Smith, 1981; Holsti, 1969; Krippendorff, 1980, 2004a, 2004b; Lombard, Snyder-Duch & Bracken, 2002, 2004; Matt, 1989; Milne & Adler, 1988; Neuendorf, 2002; Oliver, 1987; Orwin & Cordray, 1985; Stock et al., 1982; Tinsley &
Weiss, 1975, 2000) across disciplines to view interrater reliability as a serious issue whenever multiple raters or coders are used, it appears that the call has fallen on deaf ears. In their 1989 review of the marketing research literature, Perreault and Leigh noted that “there is no accepted standard for evaluating or reporting the reliability of coded data – even if multiple judges are used” (p. 137). Steiner, Lane, Dobbins, Schnur, and McConnell (1991) reviewed meta-analyses in organizational behavior and human resources management and found that only 9 of 35 studies reported interrater agreement for any coding they did. In 1998, Milne and Adler explored content analyses of social and environmental disclosures and found that there were multiple methods of assessing interrater reliability and the choice of such methods was often arbitrary. Lombard, Snyder-Duch, and Bracken (2002) analyzed 200 content analyses in the mass communications literature and found that only 69% of the content analyses contained any report of intercoder reliability.

On a positive note, Neuendorf (2002), a content analyst, has noted that “there is growing acknowledgment in the research literature that the establishment of intercoder reliability is essential, a necessary criterion for valid and useful research when human coding is employed” (p. 142). However, this current attention to intercoder reliability rises from a period where many researchers were “less than rigorous in their reliability assessment” (Neuendorf, 2002, p. 142).

Threats to Reliability in Meta-Analysis

Because the value of a meta-analysis is dependent on the reliability of the coding scheme, it is essential that meta-analysts are attentive to issues of interrater reliability.
Neuendorf (2002) has listed four key threats to interrater reliability: (a) a poorly executed coding scheme, (b) inadequate coder training, (c) coder fatigue, and (d) the presence of a rogue coder.

**Coding scheme.** Perreault and Leigh (1989) have echoed the importance of a sound coding scheme and suggest that meta-analytic approaches “can be effective only if the researcher has first developed judgment-based nominal scale coding schemes that bring structure to the diverse research studies in an area” (p. 136). Stock (1994) also has reiterated the importance of a carefully developed coding scheme: “Coding is more accurate if the coding forms and code book are designed well” (p. 129). Threats to reliability can arise from ambiguous directions in a code book, coding conventions that are not well defined, or failure to pilot the code book and coding forms and make necessary changes (Neuendorf, 2002).

**Coder training.** Once coding procedures are established, Stock (1994) has recommended that all coders involved in the study participate in training sessions as a group. The purpose of coder training is to ensure that coders “use the forms, conventions, and procedures of the synthesis in identical ways” (p. 134). Clearly, if coders are not coding in a uniform manner, intercoder reliability becomes an issue.

**Coder fatigue.** Another threat to reliability is coder fatigue or coder drift. Coder fatigue results in unstable coding over time by an individual coder. This instability typically occurs because long hours, often over a period of many months, are required to code a set of studies for meta-analysis. It is reasonable to assume that there are occasions where coding improves over time, because coders get better with practice. However, a
more potent threat to reliability occurs when a coder fails to attend to the task due to boredom or fatigue, which results in a coding accuracy that diminishes over time (Orwin, 1994).

*Rogue coders.* Neuendorf (2002) has brought a rare, but real, threat to reliability to the forefront: the rogue coder. Described as a coder “who simply cannot – or will not – be trained to achieve reliability,” the rogue coder is a potential threat to any meta-analysis.

As outlined above, the meta-analyst is not exempt from measurement issues in research. In order to ensure the credibility of his or her research, the meta-analyst must be aware of and mitigate the threats to reliability that are inherent in the data collection process.

*Intercoder Agreement versus Intercoder Reliability*

Tinsley and Weiss (2000) have suggested that many researchers incorrectly use the terms interrater reliability and interrater agreement as synonyms. In fact, Tinsley and Weiss have emphasized that indices of interrater reliability and interrrater agreement provide “fundamentally different information” (p. 97). Therefore, before embarking on a discussion of the methods of assessing intercoder reliability, it is important to address the distinction between intercoder reliability and intercoder agreement.

Under the classical true score model, the reliability coefficient is mathematically defined as the ratio of true score variance to observed score variance (Crocker & Algina, 1986). When assessing internal consistency reliability, the typical scenario is as follows: a sample of persons completes an instrument consisting of multiple items (with the
assumption that all items are measuring the same construct) at a single point in time. With each person assessed a single time, any variance in the scores on the instrument results from differences among the persons tested. Therefore, variance among people is attributed to true score variance and any nonsystematic variance is interpreted as error variance (Tinsley & Weiss, 2000).

An extension of this discussion is interrater reliability. With rating scales, the scenario is similar, except multiple raters substitute for multiple items. In this case, the assumption is that (a) raters rate the objects on an attribute that is constant at the time of measurement and (b) all raters are rating the same object. For these reasons, the variability in ratings results from differences among the rated objects. Therefore, variance among the rated objects is attributed to true score variance and any nonsystematic variance is interpreted as error variance (Tinsley & Weiss, 2000).

Because interrater reliability is calculated using correlational indices, it is not necessary for the raters to assign the rated objects exactly the same rating for interrater reliability to be high. Recall that correlational indices measuring the degree of association between two profiles are insensitive to differences in level and scatter because scores are standardized ($M = 0, SD = 1$). Therefore, interrater reliability represents the degree to which ratings for objects are identical when expressed as deviations of their means. As such, interrater reliability is sensitive to the relative ordering of the rated objects. That is, high interrater reliability indicates that the relation of one rated object to other rated objects is the same across judges, but does not indicate that absolute numbers used to express these relations are identical. For this reason, interrater reliability indices are
appropriate when the relative ordering of the rated objects is of interest (Tinsley & Weiss, 2000).

When the absolute value of the ratings or the meanings of the ratings as defined by points on a scale are of concern, indices of interrater agreement are more appropriate. Interrater agreement is an index of the extent to which multiple raters assign exactly the same rating to a particular object. Therefore, interrater agreement indices are sensitive to differences in level, scatter, or shape. When raters use a numerical scale, high interrater agreement is indicative that the raters assigned precisely the same numerical values when rating the same object (Tinsley & Weiss, 2000).

In the meta-analysis literature, use of the term *intercoder reliability* is more prevalent than the term *intercoder agreement*. Although, with respect to categorical variables, the terminology may be technically incorrect, meta-analysts often interpret the term “reliability” in the same fashion that content analysts do: reproducibility, the degree to which coding procedures produce the same coding decisions when two or more different coders analyze a text (Krippendorff, 2004a). Further, meta-analysts often utilize quantitative variables and must assess reliability accordingly. Therefore, in order to avoid a debate on semantics, this paper uses the term *intercoder reliability* with the understanding that when it comes to categorical variables, meta-analysts often measure agreement and infer reliability.
Approaches to Assessing Intercoder Reliability

Across-the-Board Approach

A common method of assessing intercoder reliability in meta-analysis is the across-the-board approach. This approach consists of computing a single agreement rate across all items on a coding form (Orwin, 1994). In their case study of reliability in meta-analysis, Stock et al. (1982) have found that most meta-analysts report intercoder reliability using the across-the-board approach, with reliabilities ranging from .70 to 1.00 (p. 11).

Although the interpretation of an across-the-board reliability is intuitively satisfying, Orwin (1994) has suggested that there are two problems with this approach: (a) it makes “little psychometric sense” because a coding sheet is a list of items thus a total-scale reliability is not meaningful and (b) it fails to bring variables which need “refining or replacement” to the forefront (p. 147). Thus, many authors (Orwin, 1994; Orwin & Corday, 1985) have recommended that intercoder reliability be calculated and reported on an item-by-item basis.

Item-by-Item Approach

An alternative to the across-the-board approach is calculating separate reliability coefficients, based on appropriate estimators, for individual coding items. This approach is much more informative than calculating reliability across-the-board (Orwin, 1994; Orwin & Corday, 1985). Further, calculating separate reliability coefficients is also a more appropriate approach because the across-the-board approach can overestimate true
intercoder agreement, especially with nominal-scale variables (Lombard, Snyder-Duch, & Braken, 2004; Orwin, 1994). This overestimation occurs because more objective items, such as those related to report identification (author, country, year, source of publication) will have very high intercoder agreement, and other, more subjective items, such as those related to study quality, may not (Orwin, 1994).

Hierarchical Approach

In the item-by-item approach, researchers typically consider the reliability estimate of each item as an independent event (Orwin 1994). However, Yeaton and Wortman (1993) have suggested that there is an inherent hierarchical structure to the reliability of variables included in meta-analyses. That is, the “lower or deeper one goes into the hierarchy, the greater the number of judgmental steps implicated in a coding decision” (Yeaton & Wortman, 1993, p. 295). As an illustrative example, Yeaton and Wortman (1993, p. 295) utilized three variables from Smith and Glass’ 1977 meta-analysis on the effects of psychotherapy:

Level 1: Condition – Is this treatment or control group data?
Level 2: Condition Type – What type of treatment or control group is used?
Level 3: Measure – For that type of condition, what measure is reported?

In this example, the levels of the hierarchy reflect the conditional nature of the coding decisions; correct coding of a variable at one level is essential for coding at all lower levels. Because of the dependent relationship between variables in a hierarchy, it follows that there is also a dependent relationship between the intercoder reliability scores of variables within different levels of the hierarchy (Yeaton & Wortman, 1993).
Yeaton and Wortman’s (1993) hierarchical approach to intercoder reliability presents another strong case for item-by-item reliability analysis of variables in a meta-analysis (Orwin, 1994). Using their hierarchical model, Yeaton and Wortman found that there is a decreasing trend in average intercoder agreement across three levels: an intercoder reliability of .93 at level 1 will result in a contingent reliability score of .80 for level 3 (p. 307). These findings suggest that using the across-the-board approach for reliability assessment can artificially inflate true agreement because of the uneven distribution of items across levels, as level 1 items are most prevalent in meta-analyses (Orwin, 1994).

Although more research is necessary to determine best practices for applying the hierarchical approach to reliability in meta-analysis, Yeaton and Wortman (1993) have made the following three recommendations: (a) increase the reliability of level 1 coding by utilizing thorough coder training procedures, (b) restrict analyses to level 1 and level 2 variables that are drawn from study-level data to reduce the systematic increase of unreliability in lower level variables, and (c) consider the hierarchical approach to reliability and report the mean reliability within each level to allow the reader to assess the validity of meta-analytic results (p. 307).

As described above, there are several approaches to assessing intercoder reliability. When it comes to actually calculating intercoder reliability of both categorical and continuous variables, the meta-analyst also has several indices of reliability to choose from. It is important to note all indices of reliability accommodate both approaches (across-the-board and item-by-item) to assessing intercoder reliability.
Assessing Intercoder Reliability: Indices of Agreement for Categorical Variables

Many variables coded for meta-analysis are categorical in nature. For example, in their meta-analysis of school-based intervention programs for aggressive behavior, Wilson, Lipsey and Derzon (2003) utilized a variable titled “study design” where coders utilized the following categories to describe the study design: randomized experiment, non-random control group designs, and single group and treatment vs. treatment (p. 140). For variables of this nature, it is reasonable and desirable for precise agreement to occur. Therefore, the appropriate index of interrater reliability for categorical variables measures the extent to which coders assign exactly the same code to a coded item (Neuendorf, 2002). The following reliability indices account for coder agreement: agreement rate, Cohen’s kappa, weighted kappa, Fleiss’ kappa, Holsti’s C.R., Scott’s pi, and Krippendorff’s alpha. To illustrate the calculation of the aforementioned coefficients of agreement, let us use the following scenario (based on data and an illustrative example from Neuendorf, 2002, p. 154-156): A meta-analyst has two coders rating ten studies. Each coder will rate all ten studies. For the purposes of this example, the coders will be categorizing the primary study into a variable based on three categories (category 1, 2, or 3). The outcome of the coding follows in Table 1.
Table 1

*Sample Data*

<table>
<thead>
<tr>
<th>Study</th>
<th>Coder A</th>
<th>Coder B</th>
<th>Agree or Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>A</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>3</td>
<td>D</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>3</td>
<td>D</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>3</td>
<td>A</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>A</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>2</td>
<td>A</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>3</td>
<td>A</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>1</td>
<td>D</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>2</td>
<td>A</td>
</tr>
</tbody>
</table>

*Agreement Rate*

Overall agreement rate (AR), also known as percent agreement, is one of the most commonly used measures of intercoder reliability in meta-analysis because it is very simple to compute (Orwin & Corday, 1985; Orwin, 1994; Stock et al., 1984). The formula for overall AR is as follows:
With the sample data from Table 1, the estimate for overall AR is:

\[
AR = \frac{7}{10} = .70
\]  

(2)

Even when the meta-analyst utilizes AR with the item-by-item approach, there are pitfalls (Cohen, 1960; Orwin, 1994; Tinsley & Weiss, 2000). With nominal-scale variables, the main problem with AR is chance agreement (Cohen, 1960; Orwin, 1994; Tinsley & Weiss, 2000). That is, a certain amount of agreement between two coders will occur by chance alone. Chance agreement is easily determined by calculating the joint probabilities of the marginals in an agreement matrix of proportions (Cohen, 1960).

*Cohen’s kappa*

To mitigate the issue of chance agreement, Cohen (1960) proposed kappa (\(\kappa\)), a coefficient of interrater agreement for nominal scales. The formula for \(\kappa\) (Cohen, 1960) is as follows:

\[
\kappa = \frac{p_o - p_c}{1 - p_c}
\]  

(3)
where, $p_o =$ the proportion of units in which the raters agreed, and $p_c =$ the proportion of units for which agreement is expected by chance, calculated using the following formula:

$$p_c = \left( \frac{1}{n^2} \right) \left( \sum pm_i \right)$$

(4)

where, $pm_i =$ each product of marginals.

To illustrate the calculation of $\kappa$ with the sample data from Table 1, it is helpful to generate a cross-tabulation table (Neuendorf, 2002, p. 154). Table 2 presents a cross-tabulation of the sample data.

<table>
<thead>
<tr>
<th>Code A</th>
<th>Category 1</th>
<th>Category 2</th>
<th>Category 3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category 1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Coder B</td>
<td>Category 2</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Category 3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>10</td>
</tr>
</tbody>
</table>

*Note.* The **bold** numbers are the numbers of studies for which the coders agree.
Using the marginals for each coder from Table 2, we can multiply and add to produce the information necessary to calculate $\kappa$. Table 3 presents the marginal values for the sample data.
Table 3

*Marginal Values for Sample Data*

<table>
<thead>
<tr>
<th>Category</th>
<th>Coder A</th>
<th>Coder B</th>
<th>A X B</th>
<th>A + B</th>
<th>Joint Marginal Proportions: $p_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>3</td>
<td>9</td>
<td>6</td>
<td>6/20 = .30</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>3</td>
<td>15</td>
<td>8</td>
<td>8/20 = .40</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>6</td>
<td>6/20 = .30</td>
</tr>
<tr>
<td>TOTAL</td>
<td>10</td>
<td>10</td>
<td>20</td>
<td>1.00</td>
<td></td>
</tr>
</tbody>
</table>
Using equation 4,

\[ p_c = \left( \frac{1}{10^2} \right) (9 + 15 + 8) = \left( \frac{1}{100} \right) (32) = .32 \]  \hspace{1cm} (5)

and substituting the value for \( p_c \) into equation 3, gives us the estimate for \( \kappa \):

\[ \kappa = \frac{.70 - .32}{1 - .32} = \frac{.38}{.68} = .56 \]  \hspace{1cm} (6)

The only assumptions required for \( \kappa \) are that (a) the measured units are independent, (b) the categories of the nominal scale are independent, mutually exclusive, and exhaustive, and (c) the judges operate independently. When obtained agreement equals chance agreement, \( \kappa = 0 \). Greater than chance agreement leads to positive values of \( \kappa \), and less than chance agreement leads to negative values. The upper limit of \( \kappa = +1.00 \), which occurs when (and only when), there is perfect agreement between the raters. The lower limit of \( \kappa \) is more complicated, since it depends on the marginal distributions. However, since researchers use \( \kappa \) as a measure of agreement, the lower limit is primarily an issue reserved for academic interest. If \( \kappa \) is less than zero (i.e., the observed agreement is less than expected by chance), it is likely to be of no further practical interest. In his description of \( \kappa \), Cohen also presented sampling characteristics and procedures for hypothesis testing and setting confidence limits (Cohen, 1960, pp. 38-44).

Tinsley and Weiss (1975) have recommended \( \kappa \) as an index of intercoder agreement whenever the same two coders rate each subject. Despite this recommendation
and the fact that $\kappa$ is a “popular and widely applied measure of reliability,” $\kappa$ is not without critics (Burton, 1981; Jones et al., 1983; Perreault & Leigh, 1989; Rust & Cooil, 1994, p. 2). Jones, Johnson, Butler, and Main (1983) have recommended using $\kappa$ with caution when raters are rating similar objects and the resultant distribution of ratings is concentrated in a small range. In this case, one or two rating categories receive a high proportion of scores while other categories receive a small or nonexistent proportion of scores. In such cases of low-variance data, Burton (1981), has maintained that $\kappa$ becomes “untrustworthy” (p. 957).

Weighted kappa

Cohen (1968) proposed the weighted kappa ($\kappa_w$) as an extension of $\kappa$. Cohen (1968) developed $\kappa_w$ for use in cases where some disagreements in assignments (i.e., some off-diagonal cells in a table of joint categorical assignment frequencies) have greater perceived gravity than others. Cohen (1968) has defined $\kappa_w$ as the proportion of weighted agreement corrected for chance, for use in situations when different kinds of disagreement are differentially weighted in the agreement index. An a priori assignment of weights to the cells in a table of joint categorical assignment frequencies accomplishes the desired weighting. Ratio scaling accomplishes the weighting, for example 6 represents twice as much disagreement as 3. When computing $\kappa_w$ using proportions, the formula is as follows:

$$
\kappa_w = 1 - \frac{\sum v_{ij} P_{oij}}{\sum v_{ij} P_{cij}}
$$

(7)
where, \( v_{ij} \) is the disagreement weight, where the \( ij \) subscript indexes the cell \((i, j = 1 \ldots k)\),

\( p_{oij} \) is the observed proportion in cell \( ij \), and \( p_{cij} \) is the chance-expected proportion in cell \( ij \), as computed for a \( \chi^2 \) contingency table.

Returning to our sample data from Table 1, let us contrive a situation where a disagreement between categories 2 and 3 is more serious than between categories 1 and 2. To do this, we will assign weights on a ratio scale using sample weights from Cohen (1968, p. 214). That is, we will assign a weight of zero to the “perfect” agreement diagonal, which will represent no disagreement. Then, we will choose a weight that represents maximum disagreement. For this example, six will indicate maximum disagreement. Although in this example we are setting weights arbitrarily, it is important to recall that, in practice, researchers must assign and justify any weighting scheme a priori.

To calculate \( \kappa_w \), it is helpful to expand the cross-tabulation table (Table 2) to include the disagreement weights, the observed cell proportions, and the chance-expected cell proportions. To calculate the chance-expected cell proportions individually, sum over the agreement diagonals and then take the product of the proportions for the row and column of the cell. Table 4 presents the expanded cross-tabulation table.
Table 4

*Expanded Cross-tabulation of Sample Data*

<table>
<thead>
<tr>
<th>Coder A</th>
<th>Category</th>
<th>Category</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>( p_i )</td>
</tr>
<tr>
<td>( (.09)^b )</td>
<td>( .20^c )</td>
<td>( .10 )</td>
<td>( .06 )</td>
</tr>
<tr>
<td>Category 1</td>
<td>0(^a)</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Coder B</td>
<td>Category 2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>( (.09) )</td>
<td>0</td>
<td>( (.15) )</td>
</tr>
<tr>
<td>Category 3</td>
<td>3</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>( (.12) )</td>
<td>( .10 )</td>
<td>( (.20) )</td>
</tr>
<tr>
<td>( p_j )</td>
<td>.30</td>
<td>.50</td>
<td>.20</td>
</tr>
</tbody>
</table>

\(^a\)Disagreement weight \((v_{ij})\). \(^b\)Chance-expected cell proportion \((p_{eij} = p_i \cdot p_j)\). \(^c\)Observed cell proportion \((p_{oij})\).

Using equation 7,

\[
\kappa_w = 1 - \frac{0(.20) + 1(.10) + 3(0) + 1(0) + \cdots + 0(.20)}{0(.09) + 1(.15) + 3(.06) + 1(.09) + \cdots + 0(.08)} = 1 - \frac{1}{2.34} = .57
\]  (8)
Although Cohen developed $\kappa_w$ after $\kappa$, it can be helpful to think of $\kappa$ as a special case of $\kappa_w$, with the same weight assigned to all disagreements. It is also important to note that researchers can develop $\kappa_w$ with cell weights that reflect agreement rather than disagreement. However, in either case (agreement of disagreement), the result is $\kappa_w$, a chance-corrected proportion of agreement. In his description of $\kappa_w$, Cohen also presented sampling characteristics and procedures for hypothesis testing and setting confidence limits (Cohen, 1968, pp. 213-218).

The use of $\kappa_w$ is only appropriate for special cases where the researcher believes that there is a relationship among the rating categories. In this situation, the researcher must assign and justify weights a priori to the rating categories based upon the theory that guides the weighting (Tinsley & Weiss, 1975). Because $\kappa$ and $\kappa_w$ are from the same family of coefficients, $\kappa_w$ is subject to the same criticisms as $\kappa$.

Fleiss’ kappa

The use of $\kappa$ and $\kappa_w$ is limited to the case where: (a) the number of raters is two and (b) the same two raters rate each subject. For this reason, Fleiss (1971) generalized $\kappa$ to the case where each of a sample of subjects is rated on a nominal scale by the same number of raters, but where the raters rating one subject are not necessarily the same as those rating another. Fleiss’ $\kappa$ reduces to Cohen’s $\kappa$ when the number of raters is two and both raters rate every subject. Fleiss’ extension of Cohen’s $\kappa$ is as follows:

$$\kappa = \frac{\overline{P} - \overline{P}_c}{1 - \overline{P}_c} \quad (9)$$
where, \( \bar{P} \), conceptualized as the overall extent of agreement, is the mean proportion of agreeing pairs out of all of the possible pairs of assignments and is calculated using the formula:

\[
\bar{P} = \frac{1}{N} \sum_{i=1}^{N} P_i
\]  

(10)

where, \( P_i \) is the proportion of agreeing pairs out of all of the possible pairs of assignments. Further, \( \bar{P}_e \), the mean proportion of agreement based on chance, calculated using the formula:

\[
\bar{P}_e = \sum_{j=1}^{k} p_j^2
\]  

(11)

where \( p_j \) is the proportion of all assignments which were to the \( j^{th} \) category, calculated using the formula:

\[
p_j = \frac{1}{Nn} \sum_{i=1}^{N} n_{ij}
\]  

(12)

finally, \( 1 - \bar{P}_e \) measures the degree of agreement attainable over and above the agreement predicted by chance. Thus, Fleiss’ generalization of \( \kappa \) is a measure of overall agreement, corrected for the amount of agreement expected by chance. In his description of \( \kappa \), Fleiss
also presents an illustrative example as well as large sample standard errors (Fleiss, 1971, pp. 378-382).

In sum, Fleiss’ $\kappa$ is appropriate when a different set of coders rate each object but the number of coders rating each object remains constant. When using Fleiss’ $\kappa$ as a measure of intercoder agreement, it is not necessary for all coders to rate each object (Tinsley & Weiss, 1975). Therefore, Fleiss’ $\kappa$ offers the meta-analyst more flexibility with respect to the use of coders than either $\kappa$ or $\kappa_w$.

Holsti’s C.R.

In the content analysis literature, Holsti (1969) has described a coefficient of reliability (C.R.), which is the ratio of coding agreements to the total number of coding decisions:

$$C.R. = \frac{2M}{N_1 + N_2} \quad (13)$$

where, $M$ is the number of coding decisions on which the two coders are in agreement, and $N_1$ and $N_2$ refer to the number of coding decisions made by coders 1 and 2, respectively. It is important to note that Holsti did not develop C.R.; he simply took Osgood’s (1959) coefficient and named it “C.R.” (Krippendorff, 2004a).

Holsti (1969) used Osgood’s coefficient to mitigate pitfalls of using frequency tabulations to assess intercoder agreement. Holsti (1969, p. 139-140) used an illustrative example to show that identical frequency tabulations of coder responses do not necessarily indicate a high level of agreement. Mathematically, Holsti’s C.R. is simply a
variation of the formula for AR. In fact, Holsti’s C.R. equals AR when the number of
coding decisions made by each coder is equal. This fact is illustrated by the sample data
from Table 1. With two coders each making 10 decisions, Holsti’s C.R. is equal to .7.
This value is identical to the AR calculated in equation 2. For this reason, Holsti’s C.R. is
subject to the same basic criticism as AR: both indices of agreement fail to account for
the extent of intercoder agreement which results by chance alone (Krippendorff, 2004a).
In his text, *Content Analysis for the Social Sciences and Humanities*, Holsti (1969) has
acknowledged the fact that C.R. fails to account for chance agreement and offers Scott’s
pi (π) as a “useful alternative” (p. 141).

*Scott’s pi*

Scott (1955) proposed π in the content analysis literature for the same reason
Cohen (1960) proposed κ in the psychology literature: to create a coefficient of
agreement which corrects for chance agreement. In fact, note that π (equation 14) and κ
(equation 3) are identical in form. The coefficients differ in how Scott (1955) and Cohen
(1960) have defined chance agreement. With π, the assumption is that the distribution of
the proportion of ratings over the categories is known and equal for all raters. By
contrast, κ does not require this assumption; instead, it allows raters to distribute their
ratings differently across categories (Tinsley & Weiss, 1975). The formula for π (Scott,
1955) is as follows:

\[
\pi = \frac{P_o - P_e}{1 - P_e} \tag{14}
\]
where, $P_o$ (observed percent agreement) represents the percentage of judgments on which two analysts independently agree upon when coding the same data, and $P_e$ is the percent agreement expected.

$P_e$ is dependent upon the number of categories in the dimension as well as the frequency with which coders utilize the categories. Therefore, assuming that the categories are mutually exclusive and the two coders’ probabilities of using any one of the categories are equal, to calculate $P_e$, the proportion of items falling into each category of a category set are squared and then summed:

$$P_e = \sum_{i=1}^{k} p_i^2$$

(15)

where $k$ is the total number of categories, and $p_i$ is the proportion of the entire sample which falls into the $i^{th}$ category. Using Table 3, the marginal values for the sample data, we first calculate $P_e$ (equation 15):

$$P_e = (.30)^2 + (.40)^2 + (.30)^2 = .09 + .16 + .09 = .34$$

(16)

Note how chance (or expected) agreement ($P_e$) as calculated for $\pi$ in equation 16 differs from how chance (or expected) agreement ($p_e$) is calculated for $\kappa$ in equation 5.

We then substitute into equation 14 to calculate $\pi$: 
Thus, $\pi$ is the ratio of the actual difference between obtained and chance agreement to the maximum difference between obtained and chance agreement; interpreted as the extent to which coding reliability exceeds chance. In his description of $\pi$, Scott also presents sampling characteristics and procedures for hypothesis testing (Scott, 1955, pp. 379-381).

Widely used in the content analysis literature, $\pi$ is appropriate for nominal level variables and two coders (Lombard, Snyder-Duch & Bracken, 2002). A criticism (Hsu & Field, 2003) of $\pi$ is that in the presence of marginal homogeneity, $\pi$ yields the same information as $\kappa$. Further, Hsu and Field (2003) have attested that the assumption of marginal heterogeneity (associated with $\kappa$) may be more realistic than the assumption of marginal homogeneity (associated with $\pi$).

*Krippendorff’s alpha*

Krippendorff’s (1980) alpha ($\alpha$) is a coefficient of agreement designed specifically for use in content analysis. Krippendorff (2004) has described $\alpha$ as not a single coefficient of agreement, but a “family” of coefficients that can accommodate multiple coders; nominal, ordinal, interval, ratio, and other metrics; missing data; and small sample sizes (p. 428). Conceptually, Krippendorff (1980) has defined $\alpha$ as:

$$\alpha = 1 - \frac{D_o}{D_e}$$  \hspace{1cm} (18)
where $D_O$ is the observed disagreement and $D_E$ is the expected disagreement. With two coders, nominal scale variables, and large samples, $\alpha$ is equal to $\pi$ (Krippendorff, 1980). To calculate $\alpha$ for nominal data with two coders, Neuendorf (2002) offers the following formula:

$$\alpha = 1 - \frac{nm - 1}{m - 1} \left( \frac{\sum pfu}{\sum pmt} \right)$$

(19)

where, $pfu$ represents the product of any frequencies for a given unit that are different (i.e., show disagreement), $pmt$ represents each product of total marginals, $n$ is number of units coded in common by coders, and $m$ is the number of coders. To calculate $\alpha$ using the sample data from Table 1, it is helpful to reconfigure the data in another table (Table 5) that shows the frequencies for both coders:
Table 5

*Frequencies for Both Coders*

<table>
<thead>
<tr>
<th>Study</th>
<th>Category 1</th>
<th>Category 2</th>
<th>Category 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>6</td>
<td>8</td>
<td>6</td>
</tr>
</tbody>
</table>

To calculate α, first calculate \( pfu \) (the product of any frequencies for a given unit that show disagreement):

\[
 pfu = (1 \times 1) + (1 \times 1) + (1 \times 1) = 3 \]  \hspace{1cm} (20)

In this case, \( pfu \) represents the disagreements for studies 3, 4, and 9. Then, calculate \( pmt \) (each product of total marginals):
Substituting in for equation 19 yields $\alpha$:

$$\alpha = 1 - \frac{(10)(2) - 1}{2 - 1} \left( \frac{3}{132} \right) = 1 - 19(.023) = 1 - .43 = .57$$

Because of its flexibility, $\alpha$ is a highly attractive coefficient of agreement. However, in practice, researchers rarely use $\alpha$ because of the extensive and tedious calculation procedure as well as the lack of “easy-to-get, easy-to-use” computer software to calculate the family of Krippendorff’s $\alpha$ indices (Lombard, Snyder-Duch, & Bracken, 2004a, p. 436; Neuendorf, 2002).

**Assessing Intercoder Reliability: Correlational Indices for Quantitative Variables**

Indices of agreement are not appropriate for quantitative variables. When interval or ratio level variables are measured, indices of correlation or covariation are more appropriate because precise agreement among coders is not likely (Neuendorf, 2002). Although there are few examples in meta-analysis where correlation alone would be a sufficient measure of intercoder reliability, the following example serves as a potential case for illustrative purposes only. In their meta-analysis of school-based intervention programs for aggressive behavior, Wilson, Lipsey, and Derzon (2003) utilized a variable titled “total hours of contact” (p. 140). For a variable of this nature, coder agreement of the precise value assigned to the variable may not be expected; therefore, indices of agreement may not appropriate. Instead, covariation of the coder ratings (when coder A
assigns a large number of total hours of contact, coder B also assigns a large number of
total of hours of contact and vice versa) is a more appropriate index of interrater
reliability. Note that when correlation indices are used, the actual agreement rate could be
0%, but the pattern of agreement, the covariation, could display an acceptable level of
reliability (Neuendorf, 2002). Intercoder reliability coefficients that consider covariation
include Pearson’s $r$ and the intraclass correlation coefficient.

*Pearson’s $r$*

When coded variables are quantitative, one measure of intercoder reliability is the
Pearson product-moment correlation coefficient (Pearson’s $r$). With respect to intercoder
reliability, Pearson’s $r$ is a measure of the degree or strength of the relationship between
the scores assigned by two coders. The formula for Pearson’s $r$ (as cited in Howell, 1999)
is as follows:

$$r = \frac{\text{cov}_{xy}}{s_x s_y} \quad (23)$$

where, $\text{cov}_{xy}$ represents the covariance of variables $x$ and $y$ and $s_x$ and $s_y$ are the standard
deviations of the two variables, $x$ any $y$, respectively.

The use of Pearson’s $r$ as a proxy for intercoder reliability results in an estimate of
the degree to which the scores produced by each coder covary. This is because the
formula for Pearson’s $r$ standardizes the coders’ scores. Therefore, Pearson’s $r$ does not
measure the degree to which the coders’ scores are in absolute agreement. Thus, coders
can produce strong, positive correlations without actually agreeing on any scores. It
follows that, an increase in absolute agreement can actually attenuate Pearson’s $r$ in situations where coders present a consistent disagreement pattern (e.g., when one coder consistently rates one item higher than another coder). This property of Pearson’s $r$ can be problematic when using it to diagnose coder consistency or reliability (Neuendorf, 2002; Orwin, 1994).

Further, when computing Pearson’s $r$, the formula always omits between-raters variance. For this reason, Ebel (1951) has recommended using the intraclass correlation coefficient as in index of intercoder reliability because researchers can choose whether to include the between-raters variance in the error term.

*Intraclass Correlation Coefficient*

Like $\kappa$ and Krippendorff’s $\alpha$, the intraclass correlation coefficient is not a single index of interrater reliability, but a family of reliability coefficients. The family of intraclass correlation coefficients, commonly used in behavioral measurement, psychometrics, and behavioral genetics, is well documented and described by McGraw and Wong (1996). The following section (Assessing Intercoder Reliability: Generalizability Theory Framework) will detail relevant formulas for intraclass correlation coefficients. Ebel (1951) has maintained that intraclass correlation coefficients are useful for estimating the reliability of ratings for three reasons: (a) researchers can choose whether to include or exclude the between-raters variance as part of the error variance dependent on the circumstances of the research; (b) there is a convenient method for estimating the precision of intraclass correlation coefficients; and
(c) intraclass correlation coefficients are calculated using analysis of variance (ANOVA), a routine computational procedure for many researchers.

Choosing to include or exclude the between-raters variance in the error term when estimating the reliability of ratings depends on the utilization of the ratings in grading, classification, or selection. Ebel (1951) has recommended removing the between-raters variance from the error term (and including this variance component in the calculation) when researchers make final decisions based upon (a) ratings consisting of averages of complete sets of ratings from all observers, (b) ratings equated from rater-to-rater as in ranks or z scores, and (c) situations where only experimental comparisons (as opposed to practical comparisons) between the rated objects are made. In situations where the researcher makes decisions by comparing single “raw” scores assigned to objects by different raters or when researchers are comparing average ratings from different groups of raters, it is essential that the between-rater variance be included as part of the error terms (Ebel, 1951).

Although the family of intraclass correlation coefficients is attractive because of the flexibility afforded by the approach, it is not without critics (Jones et al., 1983; Selvage, 1976). Because intraclass correlation coefficients are ANOVA-based, Selvage (1976) has suggested interpreting coefficients with caution when gross violations of the assumption of normality occur. Further, Selvage (1976) has offered the gentle reminder that the ANOVA-based approach requires substantial between items variance to produce a substantial indication of agreement. These concerns may or may not be relevant to the calculation of interrater reliability as significance testing is not generally associated with such calculations. Further, Jones et al. (1983) have cautioned that intraclass correlation
coefficients are not appropriate when the distribution of the ratings is concentrated in a small range.

Confusion can also arise when researchers are selecting the appropriate intraclass correlation coefficient for their work. McGraw and Wong (1996) have described 10 intraclass correlation coefficients and have offered a flow chart to assist researchers in selecting an appropriate intraclass correlation coefficient. Although this may seem to be a disadvantage of using an intraclass correlation coefficient to estimate interrater reliability, Orwin (1994) has suggested that the different variants of the intraclass correlation coefficient “reinforce the idea that a good interrater reliability assessment requires the synthesist to think about the appropriate interrater reliability design, not just calculate reliability statistics” (p. 151). Because of the close relationship between intraclass correlation coefficients and generalizability coefficients used in generalizability theory, the generalizability theory framework is often useful in designing interrater reliability studies (Orwin, 1994).

Assessing Intercoder Reliability: Generalizability Theory Framework

Oftentimes, it is best to conceptualize the family of intraclass correlation coefficients and their relationship to intercoder reliability through generalizability theory. In fact, Crocker and Algina (1986) have suggested that “the most flexible and useful approach for estimating interrater reliability coefficients is through the application of generalizability” (p. 143). Generalizability theory uses ANOVA to estimate reliability coefficients and error variance (Crocker & Algina, 1986). Because generalizability theory with a single facet encompasses four designs and is useful for both absolute (criterion-
referenced) and relative (norm-referenced) decisions, it is especially useful to the meta-
analyist who must estimate intercoder reliability under a variety of situations for both
categorical and quantitative variables.

The four basic designs for generalizability studies are:

1. A single rater assesses each object; the same rater assesses all objects.
2. Several raters assess each object; all raters assess all objects.
3. A single rater assesses each object; a different rater assesses each object.
4. Several raters assess each object; different raters assess each object

(Crocker & Algina, 1986).

Designs 1 and 2 are common in meta-analysis (Orwin, 1984). In design 1, a single coder
rates each study and the same coder rates all studies. Typically with design 2, a pair of
coders rates each study and the same pair of coders rates all studies. Because the focus of
this work is on interrater reliability, the illustrative examples that follow will utilize
design 2.

*Generalizability Theory for Relative Decisions vs. Absolute Decisions*

When the meta-analyst is assessing scale-type measures, a norm-referenced
approach to the generalizability study is appropriate. With the norm-referenced approach,
patterns of agreement, rather than absolute agreement, are the focus of the reliability
study. The norm-referenced approach examines coders’ scores based upon how they
compare or relate to each other. Conversely, with categorical variables, the criterion-
referenced approach is most appropriate. With the criterion-referenced approach, absolute
agreement is the focus of the reliability study. An absolute decision is not dependent on
patterns of agreement but by exact agreement among coders on the value of a specific variable. In generalizability theory, the generalizability coefficients for absolute and relative decisions differ with respect to the error term. In criterion-referenced designs, coder variance always matters, and therefore, it is included in the error term (Crocker & Algina, 1986).

A Generalizability Coefficient for a Design 2 Study with a Norm-Referenced Approach

As an illustrative example (based on Johanson, 2003), let us begin by creating a scenario typical to meta-analysis: A meta-analyst has two coders rating three studies. Each coder will rate all three studies. This, by definition, is a generalizability design 2. For the purposes of this example, the coders will be rating each study using a 10-point scale. The meta-analyst has decided that absolute agreement between the raters is not the priority; rather covariation, or consistency, of the coder ratings is a more appropriate proxy of reliability. Therefore, the meta-analyst is taking a norm-referenced approach. The coders’ paired scores for each study (data from McGraw & Wong, 1996) are as follows: (2,4), (4,6), and (6,8).

The first step in a generalizability study is to choose the appropriate form of ANOVA to analyze the coders’ scores. In this example, the first systematic source of variance is associated with differences among the studies (the study factor). In generalizability theory, the ANOVA model always treats the study factor as a random factor. Because there are two coders rating each of the three studies, the second source of systematic variance in the ANOVA model is the variability among coders. Thus, the coders become the second factor (the coder factor) in the ANOVA model. Because both
coders are rating all three studies, the coder factor is a fixed factor. Thus, the appropriate ANOVA model for this example is a two-way mixed model with measures of consistency. It is important to note that when using a mixed model (study factor fixed and coder factor random), inferences related to intercoder reliability are limited to the particular pair of raters used in the study (Nichols, 1999).

Table 6 presents the ANOVA table resulting from the sample data.

Table 6

Analysis of Variance for Coder Scores

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Studies</td>
<td>2</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>Within Studies</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between Coders</td>
<td>1</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Residual</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>3</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>5</td>
<td>22</td>
<td>4.4</td>
</tr>
</tbody>
</table>

The formulae for the generalizability coefficients are easiest to understand and manipulate in the form of study, coder, and residual variances. Therefore, we will convert the mean squares from the ANOVA table into their corresponding variances. In general, the formula for the study variance is:

$$
\sigma^2_{study} = \frac{MS_{study} - MS_{residual}}{n_{coder}}
$$  (24)
where \( MS \) denotes mean square and \( n \) denotes number. The general formula for the coder variance is:

\[
\sigma_{\text{coder}}^2 = \frac{MS_{\text{coder}} - MS_{\text{residual}}}{n_{\text{study}}} \quad (25)
\]

Substituting in the values from the ANOVA table (Table 6) results in:

\[
\sigma_{\text{study}}^2 = \frac{8 - 0}{2} = 4 \quad (26)
\]

and

\[
\sigma_{\text{coder}}^2 = \frac{6 - 0}{3} = 2 \quad (27)
\]

and, of course

\[
\sigma_{\text{error}}^2 = MS_{\text{residual}} = 0 \quad (28)
\]

For design 2, the formula for the generalizability coefficient with a normative (norm-referenced) decision (\( \rho_i^2 \)) is as follows:
\[
\rho_i^2 = \frac{\sigma_{\text{study}}^2}{\sigma_{\text{study}}^2 + \left( \frac{\sigma_{\text{error}}^2}{n_{\text{coder}}} \right)}
\]  

(29)

With these data, the estimate for \( \rho_i^2 \) is:

\[
\rho_i^2 = \frac{4}{4 + \frac{0}{2}} = \frac{4}{4 + 0} = 1
\]  

(30)

At this point, the merging of generalizability theory and the intraclass correlation become clear. The formula for the design 2 generalizability coefficient (equation 29) is the same as that of McGraw and Wong’s (1996) intraclass correlation coefficient C,k [ICC(C,k)], which is the appropriate coefficient for a two-way mixed effects (column effects fixed and row effects random) model. ICC(C,k) estimates the “degree of consistency for averages of \( k \) independent measures under the fixed levels of the column factor” (McGraw & Wong, 1996, p. 35).

In this example, ICC(C,k) is equal to 1.00. For these data, Pearson’s \( r \) is also equal to 1.00. This coincidence could leave the reader curious about the relationship between Pearson’s \( r \) and ICC(C,k). The key theoretical difference between Pearson’s \( r \) and ICC(C,k) is that Pearson’s \( r \) is a linearity index and ICC(C,k) is an additivity index. That is, Pearson’s \( r \) is a measure of the degree to which one variable \( y \) equates to another variable \( x \) using a linear transformation \( (y = ax + b) \). ICC(C,k) is an additivity index because it measures the degree to which one variable \( y \) equates to another variable \( x \) by
adding a constant \( y = x + b \). The resulting practical distinction is that differences in the sample variances for variables \( x \) and \( y \) attenuate ICC(C,k) relative to Pearson’s \( r \).

Intraclass correlations are based the assumption of equal variance. In the case of our example, variance differences between coders indicate a lack of agreement between the observations. Therefore, ICC(C,k) is sensitive to coder disagreement. Using the following pairs of coder scores from McGraw and Wong (1996), (0,4), (5,5), (10, 6), the attenuation of ICC(C,k) as compared to Pearson’s \( r \) is clearly illustrated. For these scores, Pearson’s \( r \) is equal to 1, which indicates perfect agreement. However ICC(C,k) for the same scores equals 0.56, an attenuated value of Pearson’s \( r \), indicating imperfect agreement (McGraw & Wong, 1996, p. 36-37).

A Generalizability Coefficient for a Design 2 Study with a Criterion-Referenced Approach

To illustrate the generalizability coefficient for a design 2 study with a criterion-referenced approach, we will return to our previous example with McGraw and Wong’s (1996) paired scores of (2,4), (4,6), and (6,8). However, in this case, the meta-analyst has decided that absolute agreement between the raters is the priority. Therefore, the meta-analyst is taking a criterion-referenced approach. Because the general design of the reliability study is the same, the appropriate ANOVA model remains a two-way mixed effects model. However, because the study has a criterion-referenced approach, ANOVA model will include coder variance as part of the overall error.
Using the ANOVA table from Table 6 and equations 24 - 28, the generalizability coefficient for a design 2 generalizability study with an absolute (criterion-referenced) decision (\( \rho^2_I \)) is:

\[
\rho^2_I = \frac{\sigma^2_{\text{study}}}{\sigma^2_{\text{study}} + \left( \frac{\sigma^2_{\text{coder}} + \sigma^2_{\text{error}}}{n_{\text{coder}}} \right)}
\]  \hspace{1cm} (31)

Note that the difference between the formula for the design 2 generalizability coefficient with a comparative (norm-referenced) decision (equation 29) and the current formula is the error variance. With absolute decisions, the coder variance is important and is accounted for by the error term.

Substituting the data from the scenario, the estimate for the design 2 generalizability coefficient with an absolute (criterion-referenced) decision is

\[
\rho^2_I = \frac{4}{4 + \left( \frac{2 + 0}{2} \right)} = \frac{4}{4 + 1} = \frac{4}{5} = .80
\]  \hspace{1cm} (32)

Recall that the design 2 generalizability coefficient with a normative (norm-referenced) decision was 1.00 for the same data. Clearly, adding the coder variance in the error term attenuates the generalizability coefficient and accounts for the lack of absolute agreement in the coders’ decisions.
The formula for the design 2 generalizability coefficient (equation 31) with an absolute (criterion-referenced) decision is analogous to McGraw and Wong’s (1996) intraclass correlation coefficient $\text{A,k}[\text{ICC(A,k)}]$. $\text{ICC(A,k)}$ is interpreted as the degree of “absolute agreement for measurements that are based on $k$ independent measurements made under the fixed levels of the column factor” (McGraw & Wong, 1996, p. 35).

Because of the flexibility to include both criterion- and norm-referenced decisions and the ease of calculation, generalizability theory and corresponding intraclass correlation coefficients show practical promise when assessing intercoder reliability. Additionally, generalizability studies can include multiple facets (e.g., types of study, gender of coder, etc.) as needed. In 1990, Hughes and Garrett suggested a generalizability theory framework for estimating intercoder reliability for quantitative data in marketing research. However, the literature remains sparse concerning this methodological innovation.

Choosing Indices of Intercoder Reliability

This review of literature has provided an overview of several indices commonly used to estimate interrater reliability. Unfortunately, there are no “universal rules of thumb or universal minimum standards” that apply to selecting and using indices of interrater reliability in applied research (Milne & Adler, 1998, p. 14). Jones, Johnson, Butler, and Main (1983) have recommended that individual researchers establish the appropriateness of chosen reliability coefficients on the basis of: (a) the nature and distribution of the data, (b) the information sought, and (c) the appropriate level of analysis for the final scores (p. 517). Perhaps Goodman and Kruskal (1954) have offered
the soundest guidance with respect to choosing an index of intercoder reliability: reliability coefficients: “measures of association should not be taken blindly from the handiest statistics textbook, but rather should be carefully constructed in a manner appropriate to the problem at hand” (p. 763).

Determining an “Acceptable” Level of Reliability

Just as there are no steadfast guidelines for selecting an appropriate reliability coefficient, there are no “established standards” for determining what signifies an acceptable level of reliability (Lombard, Snyder-Duch, & Bracken, 2002, p. 593). Landis and Koch (1977) have offered nomenclature for describing the relative strength of agreement associated with the $\kappa$ family of coefficients, with values between .81 and 1.00 signifying an “almost perfect agreement” and values between .61 and .80 indicating “substantial agreement” (p. 165). It is important to note that Landis and Koch (1977) have clearly admitted that the divisions are “clearly arbitrary” (p. 165). Krippendorff (2004b) has suggested that researchers rely only on variables with reliabilities above Krippendorff’s $\alpha = .80$ and consider variables with reliabilities between $\alpha = .667$ and $\alpha = .80$ cautiously (p. 241). After reviewing the work on interrater reliability, Neuendorf (2002) has proposed the following conventional rule: “reliability coefficients of .90 or greater would be acceptable to all, .80 or greater would be acceptable in most situations, and below that, there exists great disagreement” (p. 143).

An advantage of utilizing the family of intraclass correlation coefficients is that confidence intervals and test statistics exist for each of the intraclass correlation coefficients defined for two-way models (McGraw & Wong, 1996). With test statistics
available, researchers can test the null hypothesis that the population value of the intraclass correlation coefficient (ρ) is a given value. McGraw and Wong (1996) have noted that tests of the null hypothesis that ρ = 0 are often not informative and suggest it is more useful to determine if ρ exceeds the small, medium, or large effect size criteria proposed by Cohen. However, it is always prudent for researchers to be mindful of the fact that, at times, “there is a difference between inferential statistical significance and substantive significance” and thus statistical significance should be interpreted cautiously (Neuendorf, 2002, p. 143).

*Software Available to Calculate Intercoder Reliability*

Lombard, Snyder-Duch, and Bracken (2004b) have provided a thorough description of the software available to calculate intercoder reliability in their online resource materials. A brief overview of the software options follows.

There are four specialized software applications created specifically to calculate one or more reliability coefficients: AGREE, Krippendorff’s alpha 3.12a, Program for Reliability Assessment with Multiple Coders (PRAM), and GENOVA (GENeralized analysis Of VARiance). AGREE is Windows based software and calculates κ and an index called D2. Krippendorff’s alpha 3.12a is a beta (test) program that is Windows based and calculates the family of Krippendorff’s α coefficients. PRAM is freeware available from Skymeg Software and calculates the following coefficients: percent agreement, Scott’s π, κ, Lin’s concordance correlation coefficient, Holsti’s C. R., Spearman’s ρ, and the Pearson correlation coefficient. Note that PRAM is an alpha (test) release and the authors have warned that it contains bugs and remains a work in progress.
GENOVA is a FORTRAN computer program developed primarily for generalizability analyses. GENOVA and program documentation are available free of charge from the Iowa Testing Programs Web site (Brennan, 2001).

Two general-purpose statistical software applications include reliability coefficients. Simstat, available from Provalis Research, can calculate: percent agreement, Scott’s \(\pi\), \(\kappa\), free marginal (nominal), Krippendorf’s r-bar, Krippendorf’s R (for ordinal data) and free marginal (ordinal). The Statistical Package for Social Sciences (SPSS) can calculate \(\kappa\) as well as intraclass correlation coefficients (Lombard, Snyder-Duch, and Bracken, 2004b; Nichols, 1999). Nichols (1999) provides detailed directions for calculating intraclass correlations (and thus, generalizability coefficients) using SPSS.

A few researchers have written macros for use with existing software packages to calculate intercoder reliability. A SAS macro is available to calculate Krippendorff’s \(\alpha\) for nominal, ordinal, and interval data, \(\kappa\), Perrault’s index, and Bennet’s index. SPSS macros are available to calculate percent agreement and Krippendorff’s \(\alpha\) for one or more two category nominal level variables with two coders (Lombard, Snyder-Duch, and Bracken, 2004b).

**Summary and Conclusions**

The meta-analyst is not exempt from addressing measurement issues. Researchers tend to agree that meta-analysts should “provide a rigorous assessment of interrater reliability” but there are no steadfast and standard guidelines pertaining to how to select a reliability coefficient or what determines an acceptable level of reliability (Bullock &
Svyantek, 1985, p. 114). Further, the question of interrater reliability is not unique to meta-analysis, as content analysts are also wrestling with related issues. Therefore, the purpose of this work is two fold: First, it will describe the state of intercoder reliability in meta-analyses related to education. Secondly, it hopes to offer information to improve best practices in intercoder reliability assessment that will transcend across disciplines.
CHAPTER THREE

Methodology

This study is a content analysis of meta-analyses published related to preschool through grade 12 education. The methods used to collect, analyze, and present the data that pertain to the overarching research question: How is intercoder reliability assessed and reported in published meta-analyses related to preschool through grade 12 education? are given in this chapter. This chapter consists of two main sections: research design and data analysis. The research design section contains five subsections: identification of the population, instrumentation, coding process, reliability samples, and interrater reliability analysis.

Research Design

Identification of the Population

A search of journal articles indexed in three databases identified prospective meta-analyses for this study. The study utilized the following three databases: (a) Education Resources Information Center (ERIC), a digital library of education-related resources sponsored by the Institute of Education Sciences of the U.S. Department of Education; (b) Education Abstracts, a database produced by the H. W. Wilson Company that covers over 500 core international periodicals, monographs, and yearbooks related to contemporary education issues; and (c) PsycINFO, a database produced by the American Psychological Association that provides abstracts and citations to the scholarly literature in the behavioral sciences and mental health. Journal articles were included the study
based on the following inclusion criteria: (a) journal articles that employ meta-analytic
methods and report meta-analytic results, and (b) meta-analyses covering content related
to education in preschool through grade 12. Further, journal articles that were
methodological and theoretical in nature with no reported meta-analytic results were not
eligible for inclusion the study.

An initial search for English-language journal articles in the ERIC database, for
the years 2000 through 2004 for which one of the keywords was “meta-analysis,” yielded
277 journal articles. Of these journal articles, many meta-analyses were from content
areas outside the scope of preschool through grade 12 education: behavioral health,
counseling, government policy, industrial and organizational psychology, psychiatry, and
strength training. To maintain the focus on preschool through grade 12 education, the
search was further limited by educational level to include: early childhood education
(birth through grade 3), preschool education (birth to kindergarten),
elementary/secondary education (kindergarten through grade 12), elementary education
(kindergarten through grades 6, 7, or 8), primary education (kindergarten through grade
3), intermediate grades (grades 4, 5, and 6), secondary education (grades 7 through 12),
middle schools (grades 4 through 9), junior high schools (grades 7, 8, and 9) and high
schools (grades 9 through 12). With these limitations and the inclusion criteria, the search
identified 72 prospective journal articles. From the prospective journal articles, 38
articles were included in the final sample.

An initial search for English-language journal articles in Education Abstracts, for
the years 2000 through 2004 for which one of the keywords was “meta-analysis,” yielded
376 journal articles. To maintain the focus on preschool through grade 12 education, the
principal investigator performed an abstract review to determine potential articles for inclusion. The abstract review identified 131 prospective journal articles. From these prospective journal articles, 43 articles were included in the final sample based on the inclusion criteria.

An initial search for English-language journal articles in the PsycInfo database, for the years 2000 through 2004 with the following age delimitations: neonatal (birth – 1 month), infancy (2 – 23 months), preschool age (2 – 5 years), school age (6 – 12 years), and adolescence (13 – 17 years) for which one of the keywords were “meta-analysis,” yielded 386 journal articles. To maintain the focus on preschool through grade 12 education, the search was further limited by the following PsycInfo classification codes: Educational Psychology, Educational Administration and Personnel, Curriculum and Programs and Teaching Methods, Academic Learning and Achievement, Classroom Dynamics and Student Adjustment and Attitudes, Special and Remedial Education, Gifted and Talented, and Educational/Vocational Counseling, and Student Services. With this limitation, the search identified 87 prospective journal articles. From the prospective journal articles, 19 articles were included in the final sample based on the inclusion criteria.

The initial data set for this study included 100 journal articles acquired from the three databases: ERIC, Education Abstracts, and PsychInfo. During the coding process, a coder discovered that one article failed to meet inclusion criteria. This resulted in the exclusion of the article from the study. Therefore, the overall data set included 99 journal articles. Within these journal articles, 10 articles reported results from more than one meta-analysis. Because some authors reported interrater reliability information for more
than one meta-analysis, each meta-analysis (whether the author reported interrater reliability information or not) was considered a separate case. Therefore, the final data set is comprised of 118 meta-analyses reported in 99 journal articles.

Instrumentation

Instrument development. The coding instrument for this study utilized items based on variables coded in two previous studies: Lombard, Snyder-Duch, and Bracken’s (2002) study that assessed the reporting of intercoder reliability in content analyses in the mass communications literature and Steiner et al.’s (1991) study which reviewed meta-analyses in organizational behavior and human resources management. Specifically, the items with 5-point rating scale (items 4c, 5a, and 5d) were adapted from Steiner et al. (1991). Appendix A presents the coding manual and Appendix B presents the coding sheet used for the study.

Variables coded. The content analysis of the characteristics of meta-analyses related to preschool through grade 12 education includes the following 48 variables:

1. Content area of meta-analysis.
2. Number of studies included in meta-analysis.
3. Coders and Coding Strategy
   a. Presence or absence of reporting of the number of coders.
   b. Number of coders participating in the coding process.
   c. Clarity of information presented regarding number of coders.
   d. Presence or absence of differentiation between coding of the reliability sample and coding of the actual sample.
e. Size of the reliability sample.

f. Clarity of information presented regarding size of reliability sample.

g. Proportion of primary studies included in the reliability sample.

h. Presence or absence of a description of the coders’ levels of expertise.

i. Description of coders (e.g., undergraduate research assistants, graduate research assistants, study authors, etc.)

4. Coder Training

a. Presence or absence of coder training procedures.

b. Amount and type of coder training.

c. Description of coder training. Rated on a 5-point Likert scale where “1” corresponds to “coder training is kept secret; not described at all” and “5” represents “coder training is described in enough detail that it would be easy to replicate the coder training procedure.”

d. Presence of absence of a criterion for consistency (e.g., stopping rule) to determine that coders were adequately trained.

5. Code Book

a. Description of code book. Rated on a 5-point Likert scale where “1” corresponds to “code book is kept secret; not described at all” and “5” represents “code book is described in enough detail that it would be easy to replicate the study.”
b. Presence or absence of a statement indicating that the code book is available from the author(s).

c. Presence or absence of a method to resolve coding discrepancies.

d. Description of resolution of coding discrepancies. Rated on a 5-point Likert scale where “1” corresponds to “method of resolution is kept secret; not described at all” and “5” represents “method of resolution is described in enough detail that it would be easy to replicate the study.”

6. Addressing Intercoder Reliability

a. Presence or absence of the author’s acknowledgment of intercoder reliability.

b. Presentation of intercoder reliability information in the article.

c. Number of sentences devoted to the general discussion of intercoder reliability.

d. Number of sentences devoted to the reporting of intercoder reliability analyses.

e. Presence or absence of at least one citation about intercoder reliability.

f. Number of citations related to intercoder reliability.

g. Resources the authors referencing with respect to intercoder reliability.

7. Reporting Intercoder Reliability

a. Presence or absence of a specific method for calculating intercoder
reliability

b. Specific method(s) reported to calculate intercoder reliability.

c. Reporting format for reliability statistics.

d. If reliability statistics are not reported, and multiple coders were used, presence or absence of a phrase(s) authors use to imply coder reliability.

e. Presence or absence of differentiation between effect size coding and study descriptor coding.

f. Presence or absence of an overall/average reliability value for a group or groups of variables related to effect size in the study.

g. Presence or absence of a range of reliability for a group or groups of variables related to effect size in the study.

h. Presence or absence of specific intercoder reliability values for one or more variables related to effect size in the study.

i. Presence or absence of intercoder reliability values for all variables related to effect size included in the meta-analysis.

j. Clarity of information presented regarding the reporting of intercoder reliability of effect size variables.

k. Presence or absence of an overall/average reliability value for a group or groups of variables related to study descriptors in the study.

l. Presence or absence of a range of reliability for a group or groups of variables related to study descriptors in the study.
m. Presence or absence of specific intercoder reliability values for one or more variables related to study descriptors in the study.

n. Presence or absence of intercoder reliability values for all variables related to study descriptors included in the meta-analysis.

o. Clarity of information presented regarding the reporting of intercoder reliability of study descriptor variables.

p. If effect size and study descriptor variables are not differentiated, presence or absence of an overall/average reliability value for a group or groups of variables in the study.

q. If effect size and study descriptor variables are not differentiated, presence or absence of a range of reliability for a group or groups of variables in the study.

r. If effect size and study descriptor variables are not differentiated, presence or absence of specific intercoder reliability values for one or more variables in the study.

s. If effect size and study descriptor variables are not differentiated, presence or absence of intercoder reliability values for all variables included in the meta-analysis.

t. If effect size and study descriptor variables are not differentiated, clarity of information presented regarding the reporting of intercoder reliability.

u. Presence or absence of a discussion of “acceptable” reliability values.
v. Range of reported reliability values.
w. Clarity of information presented regarding the reporting of the reliability values.
x. Lowest reliability criterion deemed “acceptable.”

**Coding Process**

For the current study, there were two coders, each coder coding all studies. The coders for this study were trained Ohio University undergraduate students. During the study period, both coders were candidates for graduation. Through coursework and tutorial experiences, the students had experience with research design and methodology as well as basic statistical knowledge. The students received course credit (PSY 390: Research in Psychology) for participating in the study.

Three coding meetings took place over a one-month period. Each meeting lasted approximately two hours. The principal investigator and both coders attended all meetings. The principal investigator facilitated all meetings. Prior to the first meeting, the coders were each given a journal article from the reliability sample, the code book, and the coding sheet via electronic mail. Coders used the code book and coding sheet to independently code the sample article.

The initial coding meeting served as a training session for the coders. The method of coder training was based on Stock’s (1994, p. 134-135) eight steps for training coders participating in a meta-analysis. At the initial coding meeting, the principal investigator presented an overview of the current study (Stock, 1994, Step 1). Using the sample study as a point of reference, the principal investigator facilitated the reading of each item on
the coding sheet and the corresponding description in the code book (Stock, 1994, Step 2). For each item, the principal investigator and coders compared the completed coding sheets from the sample study, identifying any discrepancies or ambiguities (Stock, 1994, Step 6, p. 134). As this process occurred, the principal investigator, with input from the coders, revised the code book and coding sheets as appropriate (Stock, 1994, Step 7, p. 134). Through this process, the principal investigator and the coders established consensus on the coding structure for the project (Stock, 1994, Step 8, p. 134). Prior to the conclusion of the meeting, the principal investigator then described the process for using the coding sheets, including data entry and management procedures, as well as organization of study materials (Stock, 1994, Step 3). The principal investigator also encouraged the coders to contact her should they have any questions or need further clarification.

After the first coding meeting, the principal investigator revised the code book and coding sheets. Once revisions were complete, the coders received the initial reliability protocol. (A detailed explanation of the reliability protocol for the entire study follows in the next section.) The principal investigator instructed the coders to code each article independently. The coders returned the coding sheets for the initial reliability sample (described later) to the principal investigator for reliability analysis prior to the second coding meeting.

The second coding meeting served three purposes: (a) to further identify and resolve any discrepancies or ambiguities in the code book and coding sheet, (b) to resolve any coding discrepancies between the two coders that occurred in the initial reliability sample, and (c) to address any concerns or questions raised by the coders. The principal
investigator facilitated this meeting by presenting the coding discrepancies that occurred in the initial reliability sample to the coders. The coders then discussed each discrepancy and reached consensus. If consensus could not be reached, the principal investigator assisted by directing the coders to the code book for guidance and facilitating a metacognitive discussion about how to code a particular item. After such discussion, the coders reached consensus.

At the conclusion of the second coding meeting, coders received the final sample, first half coding protocol (described later). Ten days after the second coding meeting, the coders returned the coding sheets from the final sample, first half coding protocol and picked up coding materials and articles for the final sample, second half coding protocol (described later). At this time, the principal investigator refreshed the process for using the coding sheets, including data entry and management procedures, as well as organization of study materials and addressed any questions or concerns raised by the coders (Stock, 1994, Step 3). The principal investigator also encouraged the coders to contact her should they have any questions or need further clarification.

The coders returned the coding sheets for all remaining studies to the principal investigator for reliability analysis prior to the third coding meeting. The third coding meeting served primarily to resolve any coding discrepancies between the two coders that occurred in the final reliability sample. The principal investigator facilitated this meeting by presenting the coding discrepancies that occurred in the final reliability sample to the coders. The coders then discussed each discrepancy and reached consensus. If consensus could not be reached, the principal investigator assisted by directing the coders to the
code book for guidance and facilitating a metacognitive discussion about how to code a particular item. After such discussion, the coders reached consensus.

**Reliability Samples**

After the initial literature search, 100 journal articles containing meta-analyses were determined eligible for inclusion in the study. Unfortunately, there is little guidance in the content analysis and meta-analysis literature as to the number of units to be included in the reliability assessment (Neuendorf, 2002; Orwin, 1994; Stock, 1994). Potter and Levine-Donnerstein (1999) have examined the state of reliability and validity in content analysis and reported that reliability samples ranged from 10% to 100% of the full sample. In their mass media research primer, Wimmer and Dominick (2003) have presented the rough guideline of 10% to 20% of the total sample. The meta-analysis literature (Orwin, 1994; Stock, 1994) has encouraged pilot studies of the coding sheet and code book using randomly selected studies from the full sample; however there is no guidance regarding the number or proportion of studies recommended for use in the pilot study. Because of the lack of clear direction with respect to the number of studies to include in the reliability assessment, this study uses the upper limit of Wimmer and Dominick’s (2003) rough guideline. Therefore, for this study, the initial reliability sample was composed of 20% of the 100 articles (20 articles) containing meta-analyses included in the full study, selected at random. The initial reliability sample served as a benchmark for establishing intercoder reliability for the study. A description of the procedures and standards used to establish intercoder reliability follows in the next sections.
After the coders coded the 20 articles in the initial reliability sample, 80 articles remained. Due to this large number of articles, it was unrealistic to have each coder code all of the remaining articles. For this reason, the principal investigator split the remaining articles between the coders, 40 articles each. During the initial coding meeting, the coders chose to split the final sample into half in order to help keep themselves motivated and on-task during the coding process. Therefore, the coding protocol for the remaining 40 articles was as follows: (a) coders received the first 20 articles at the second coding meeting, (b) coders returned the completed coding sheets for the first 20 articles ten days later and received the final 20 articles at that time, and (c) coders returned the completed coding sheets for the final 20 articles ten days later.

In order to assess interrater reliability over time, each half of the final sample included a 20% reliability sample that included articles that both coders coded, selected at random. Therefore, this study included four reliability samples:

1. Initial reliability analysis: This reliability analysis consisted of 20% of the 100 articles (20 articles) determined eligible for inclusion. The purpose of this analysis was to establish baseline interrater reliability.

2. Final sample, first half reliability analysis: This reliability analysis consisted of 20% of the 40 articles (8 articles) included in the coding of the first half of the final sample. The purpose of this analysis was to examine interrater reliability over time.

3. Final sample, second half reliability analysis: This reliability analysis consisted of 20% of the 40 articles (8 articles) included in the coding of the second half of the final sample. The purpose of this analysis was also
to examine interrater reliability over time.

4. Overall reliability analysis: This reliability analysis consisted of all articles (36 articles) utilized in the previous three analyses. During the final sample, second half coding process, a coder discovered that one article failed to meet inclusion criteria. This resulted in exclusion of the article from the study. Thus, the overall reliability analysis accounted for 36.4% of the total articles ($N = 99$) included in the study. The purpose of this analysis was to establish overall interrater reliability.

*Interrater Reliability Analysis*

Per the discussion in Chapter Two, there are several indices used to estimate interrater reliability and agreement. Unfortunately, the literature review revealed that there are no general guidelines or standards with respect to the selection and utilization of interrater reliability and agreement indices, and the range of values for reliability indices determine an acceptable level of reliability.

This study used an item-by-item approach for all four reliability analyses described previously: (a) initial reliability analysis; (b) final sample, first half reliability analysis; (c) final sample, second half reliability analysis; and (d) overall reliability analysis. The principal investigator calculated separate reliability coefficients, based on appropriate estimators, for individual coding items. As discussed in Chapter Two, an item-by-item approach is much more informative than calculating reliability across-the-board (Orwin & Corday, 1985; Orwin, 1994).
At the onset of the study, the principal investigator chose a generalizability theory framework to assess interrater reliability for this study because of the flexibility to include categorical and quantitative variables and accommodate criterion- and norm-referenced decisions. Further, generalizability coefficients are easy to calculate via the intraclass correlation using SPSS. The rationale for the design of the generalizability study is as follows. In this study, two coders rated all studies in the reliability samples. This, by definition, is generalizability design 2. Therefore, the appropriate ANOVA model for this study is a two-way mixed model. It follows that, for categorical variables, the appropriate generalizability coefficient is a design 2 generalizability coefficient with an absolute (criterion-referenced) decision (recall that this coefficient is the same as ICC[A,k]). For quantitative variables, the appropriate generalizability coefficient is a design 2 generalizability coefficient with a relative (norm-referenced) decision (recall that this coefficient is the same as ICC[C,k]). Appendix C provides information on how to link to Nichols’ (1999) paper, which gives a detailed explanation of how to calculate intraclass correlation coefficients (and therefore generalizability coefficients) using SPSS, via the World Wide Web.

However, after examining the distribution of the data, it became apparent that straying from the generalizability theory framework was necessary and justified. Due to its ease of interpretation, the principal investigator chose percentage agreement as the secondary index of interrater reliability. Although percentage agreement does not account for chance agreement, it was the only other index of interrater reliability that could accommodate all other variables in the study. Further, percentage agreement is easy to
calculate using SPSS, as Lombard’s (2005) SPSS macro for calculating percent agreement is available free on the World Wide Web.

Although the framework for the reliability analysis was designed prior to the collection of data, choosing a specific reliability index for each item involves examining the nature and distribution of the data to select the most appropriate index. The paragraphs that follow present the decision-making process for choosing the appropriate index of interrater reliability for all variables in the study. Chapter Four presents complete results from all of the reliability analyses conducted.

*Items 1a – 1d.* Item 1 asked the question, “What is the content area of the meta-analysis?” Coders could choose from 22 responses (see Appendices A and B), including a category marked “other” which offered the opportunity for an open-ended response. Coders could choose multiple content areas if applicable. Coders ranked their decisions with “1” indicating the most appropriate choice and subsequent choices followed suit. Because this was a multiple response item, it resulted in four variables; with the first variable, 1a, representing the coders’ first choices and so on up to variable 1d. Variables 1a to 1d were categorical, therefore the principal investigator chose the generalizability coefficient for a design 2 generalizability study with an absolute (criterion-referenced) decision (ICC[A,k]) as the index of interrater reliability.

Although ICC(A,k) was appropriate for variable 1a, it was not an appropriate estimator of interrater reliability for variables 1b to 1d. Because variables 1b to 1d represented the coders’ second, third, and fourth choices for content areas, the number of responses declined for each subsequent variable. Further, after examination of frequency distributions, the distribution of ratings was concentrated in a small range. As stated in
Chapter Two, intraclass correlation coefficients are not appropriate in such situations (viz., Jones et al., 1983). For these reasons, the principal investigator chose percentage agreement as the index of interrater reliability for variables 1b to 1d.

Item 2. Item 2 asked the question, “How many primary studies are included in the meta-analysis?” This was an open-ended item where coders simply recorded the number of primary studies. Because this resulted in a quantitative variable, the principal investigator chose the generalizability coefficient for a design 2 generalizability study with a relative (norm-referenced) decision (ICC[C,k]) as the index of interrater reliability.

Items 3a – 3i. Items 3a to 3i consisted of a series of questions exploring the coders and the coding process. Item 3a asked the question, “Did the author report the number of coders involved in coding the primary studies?” Coders responded with “yes” or “no.” This variable was categorical, therefore the principal investigator chose a generalizability coefficient for a design 2 generalizability study with an absolute (criterion-referenced) decision (ICC[A,k]) as the index of interrater reliability.

Item 3b1 asked the general question “How many coders were involved in coding the primary studies?” Coders responded with one of three choices: “one coder,” “more than one coder,” or “The author does not report any details about the person(s) involved in the coding process.” As in the case with Item 3a, the principal investigator chose the generalizability coefficient for a design 2 generalizability study with an absolute (criterion-referenced) decision (ICC[A,k]) as the index of interrater reliability.

Item 3b2 asked coders to note the specific number of coders involved in coding the primary studies. Coders responded to the open-ended question with the number of coders or if it was clear that there was more than one coder, but the exact number of
coders was not reported, coders responded with “not reported.” Because the structure of this variable had both categorical and quantitative components, the principal investigator chose percentage agreement as the index of interrater reliability.

Item 3c asked the question “Was the information needed to code variable 3b clearly stated in the text?” Coders responded with one of three choices: “yes,” “no,” or “I was unable to ascertain any information about the individuals involved in coding the primary studies.” As in the case with Items 3a and 3b1, the principal investigator chose the generalizability coefficient for a design 2 generalizability study with an absolute (criterion-referenced) decision (ICC[A,k]) as the index of interrater reliability.

Item 3d asked the question “Did the author(s) differentiate between coding of the reliability sample and coding of the actual sample?” Coders responded with either “yes” or “no.” As in the case with Items 3a, 3b1, and 3c, the principal investigator chose the generalizability coefficient for a design 2 generalizability study with an absolute (criterion-referenced) decision (ICC[A,k]) as the index of interrater reliability.

Item 3e asked coders to note the size of the reliability sample (the number or primary studies in the reliability sample). For this item, only the studies reporting a reliability sample were included in the analysis. Although this variable is quantitative in nature, it was important to the principal investigator that the coders have absolute agreement on the size of the reliability sample. Therefore the principal investigator chose the generalizability coefficient for a design 2 generalizability study with an absolute (criterion-referenced) decision (ICC[A,k]) as the index of interrater reliability. However, for this variable, the final sample second half reliability sample had a distribution of ratings concentrated in a small range; therefore, the intraclass correlation was an
appropriate measure of interrater reliability. For this reason, the principal investigator used percentage agreement as the index of interrater reliability for the final sample second half reliability sample.

Item 3f asked the question “Was the information needed to code variable 3e clearly stated in the text?” This item was similar to item 3c. Therefore, the principal investigator used the same approach to analyzing the reliability of this variable: the generalizability coefficient for a design 2 generalizability study with an absolute (criterion-referenced) decision (ICC[A,k]).

Item 3g asked coders to calculate the proportion of primary studies that were included in the reliability sample. For this variable, only the studies reporting a reliability sample were included in the analysis. Because this variable is quantitative in nature, the principal investigator chose the generalizability coefficient for a design 2 generalizability study with a relative (norm-referenced) decision (ICC[C,k]) as the index of interrater reliability. However, for the final sample second half reliability sample, the proportion of studies included in the reliability sample was the same for all three cases. Because intraclass correlation coefficients are ANOVA-based, they require between items variance in order to be calculated (Selvage, 1976). Because there was no item variance in this case, the principal investigator chose percent agreement as the index of interrater reliability for this reliability analysis.

Item 3h asked coders the question “Did the author provide a description of the coders’ levels of expertise?” Coders responded with either “yes” or “no.” As this item was very similar to item 3d, the principal investigator chose the generalizability
coefficient for a design 2 generalizability study with an absolute (criterion-referenced) decision (ICC[A,k]) as the index of interrater reliability.

Items 3i1 and 3i2 resulted from the question “How did the author describe the coders’ levels of expertise?” Coders chose from four responses: “graduate students/graduate research assistants,” “undergraduate students/undergraduate research assistants,” “study authors,” or “other,” which gave the option for coders to write in an open-ended response. With this item, coders could choose more than one response, which generated two variables. As in the case for items 1b to 1d, ICC(A,k) was not an appropriate estimator for interrater reliability due to the restriction of range associated with the distribution of ratings. As such, the principal investigator chose percentage agreement as the index of interrater reliability.

*Items 4a – 4d.* Items 4a to 4d consisted of series of questions related to coder training. Item 4a asked the question “Does the text state how coders were trained?” Coders responded with “yes” or “no.” For this variable, most of the coders responses were “no.” Because of this, ICC(A,k) was not an appropriate estimator for two reasons: the restriction of range associated with the distribution of ratings and the corresponding lack of between items variance (as in the case with item 3g). For these reasons, the principal investigator chose percentage agreement as the index of interrater reliability.

Items 4b1 to 4b4 asked the coder to note the amount and units of coder training and item 4c asked the question “How does the author describe the coder training procedure?” For item 4c, coders responded on a 5-point Likert scale (see Appendices A and B). Because study authors consistently did not report coder training, the restriction of range associated with the distribution of ratings and the corresponding lack of between
items variance (as in the case with item 4a) were primary factors in choosing percentage agreement as the most appropriate index of interrater reliability.

Item 4d1 asked the question “Does the author describe a specific reliability criterion which establishes consistency (e.g., stopping rule) to determine that coders were adequately trained?” Coders responded “yes” or “no.” Further, if coders responded “yes,” they noted the reliability criterion (item 4d2). As in the previous items (items 4b1 – 4b4 and 4c), authors consistently did not report a stopping rule, resulting in the restriction of range and lack of variance issues discussed previously. Therefore, the principal investigator chose percentage agreement as the index of interrater reliability.

 Items 5a – 5d. Items 5a to 5d consisted of a series of questions related to the code book. Item 5a asked the question “How does the author describe the code book?” Coders responded to this item using a 5-point rating scale (see Appendices A and B). Because the item used a 5-point rating scale, the principal investigator had to choose between an absolute (criterion-referenced) decision and a consistency (norm-referenced) decision. Ultimately, the principal investigator determined that the focus of this reliability analysis was on measuring absolute agreement between the coders, rather than measuring decision consistency. Absolute agreement between coders is essential when the primary focus of the study is descriptive in nature. For this reason, the principal investigator chose the generalizability coefficient for a design 2 generalizability study with an absolute (criterion-referenced) decision (ICC[A,k]) as the index of interrater reliability.

Items 5b and 5c were similar in nature. Item 5b asked the question “Do the authors state that a copy of the code book (or coding protocol) is available upon request?” Item 5c asked coders to report if the text stated how coding discrepancies were resolved.
For both items, coders responded with “yes” or “no.” As in the case with items 4a – 4d, coder responses to items 5b and 5c resulted in a situation where restriction of range and lack of item variance caused the use of ICC(A,k) to be inappropriate. Therefore, the principal investigator chose percentage agreement as the index of interrater reliability.

Item 5d was similar to item 5a. Item 5d asked the question “How does the author describe the resolution of coding discrepancies?” Coders responded to this item using a 5-point Likert scale (See Appendices A and B.). Using the same reasoning described for item 5a, the principal investigator chose the generalizability coefficient for a design 2 generalizability study with an absolute (criterion-referenced) decision (ICC[A,k]) for the initial reliability sample and the final sample first half reliability analyses. However, for the final sample second half reliability analyses, the lack of between items variance made the use of ICC(A,k) inappropriate. Further, lack of between items variance in the final sample second half data made it inappropriate to use ICC(A,k) for the overall reliability analysis. Therefore, the principal investigator chose to percentage agreement as the index of interrater reliability for the final sample second half and overall reliability analyses.

*Items 6a – 6f.* Items 6a to 6f consisted of a series of questions related to how authors addressed intercoder reliability in their articles. Item 6a asked coders “Is reliability acknowledged in the article?” Coders responded “yes” or “no.” For this item, the distribution of ratings was not restricted, so principal investigator chose the generalizability coefficient for a design 2 generalizability study with an absolute (criterion-referenced) decision (ICC[A,k]) as the index of interrater reliability.

Item 6b asked the question “How did the study author(s) present reliability information in the text? Coders could choose from four responses, “Separate section with
‘reliability’ heading in text,” “One or more complete paragraphs about reliability in text,” “Less that a paragraph about reliability in text,” or “Reliability discussed in footnote(s) only.” Similar to items 4a – 4d, over 85% of the time, coders chose the response “Less than a paragraph about reliability in text,” resulting in a lack of between items variance. For this reason, the principal investigator chose percentage agreement as the index of reliability.

Items 6c and 6d were very similar. Item 6c asked coders to note the number of sentences in the text that were devoted to the general discussion of reliability and item 6d asked coders to note the number of sentences that were devoted to the reporting of reliability analyses. Because this resulted in a quantitative variable, the principal investigator calculated the generalizability coefficient for the design 2 generalizability study with a relative (norm-referenced) decision (ICC[C,k]) as the index of interrater reliability.

Item 6e asked the question “Does the article include at least one citation (reference) about intercoder reliability?” Coders responded “yes” or “no.” As a follow-up question, item 6f asked coders to note the number of citations or references related to intercoder reliability. In general, authors consistently did not include any citations related to intercoder reliability, resulting in the restriction of range and lack of variance issues discussed previously. Therefore, the principal investigator chose percentage agreement as the index of interrater reliability.

*Items 7a – 7x.* Items 7a – 7x consisted of a series of questions related to the reporting of intercoder reliability. Item 7a asked the question “Does the author report a specific method for calculating intercoder reliability?” Coders responded “yes” or “no.”
For this item, the distribution of ratings and between items variance were not restricted for the final sample first and second half reliability analyses, so principal investigator chose the generalizability coefficient for a design 2 generalizability study with an absolute (criterion-referenced) decision (ICC[A,k]) for those reliability analyses. However, for the initial reliability sample, there was insufficient between items variance (as was the case for items 4a – 4d), which resulted in ICC(A,k) being inappropriate for the initial sample reliability analysis. Consequently, the addition of the data to the overall reliability sample caused insufficient between items variance in the overall reliability analysis. Therefore, the principal investigator chose to use percentage agreement as the index of interrater reliability for the initial and overall reliability analyses.

Items 7b1 to 7b3 were similar to the multiple response scenario described with item 1. Item 7b asked the question “What method was used to report intercoder reliability?” Coders could choose from 15 responses (see Appendices A and B), including a category marked “other” which offered the opportunity for an open-ended response. Coders could choose multiple methods if applicable. Because this was a multiple response item, it resulted in three variables, with the first variable, 7b1, representing the coders’ first choices and so on up to 7b3. This variable was categorical, therefore the principal investigator chose the generalizability coefficient for a design 2 generalizability study with an absolute (criterion-referenced) decision (ICC[A,k]).

However, ICC(A,k) was not an appropriate estimator for interrater reliability for variables 7b2 and 7b3. Because variables 7b2 and 7b3 represented the coders’ second and third choices for methods, the number of responses declined for each subsequent variable. Further, after examination of frequency distributions, the distribution of ratings was
concentrated in a small range. Therefore, the principal investigator chose percentage agreement as the index of interrater reliability for variables 7b2 and 7b3.

Item 7c asked the question “How did the author(s) report reliability statistics?” Coders chose from four responses: “only in a text description,” “only in a table(s),” “a combination of text description and table(s),” or “the authors did not report reliability statistics.” For this item, restriction of range and between items variance did not appear to be a problem, so the principal investigator chose the generalizability coefficient for a design 2 generalizability study with an absolute (criterion-referenced) decision (ICC[2,k]) as the index of interrater reliability.

Item 7d was a variable that was qualitative in nature; therefore, quantitative reliability analyses were not appropriate for this item. Item 7e asked the question “Did the author differentiate between effect size coding and study descriptor coding?” Coders responded “yes” or “no.” For this item, coders responded “no” over 80% of the time, resulting in a lack of between items variance. For this reason, the principal investigator chose percentage agreement as the index of interrater reliability.

Items 7f – 7t consisted of a series of items that asked coders to determine what information related to reliability analyses study authors were reporting. For each item, coders responded “yes,” “no,” or “standardized reliability criterion values not reported.” Reliability analyses for items 7f – 7o included cases where coders designated that study authors differentiated between effect size coding and study descriptor coding (n = 9). Items 7p – 7t included cases where coders designated that the study authors did not differentiate between effect size coding and study descriptor coding (n = 38). Note that the sum of these sample sizes does not equal the total sample size (n = 42) because there
was some coder discrepancy as to whether or not authors differentiated between effect size coding and study descriptor coding.

In most cases, authors did not report reliability analyses. This resulted in a lack of between items variance for items 7f - 7t. For this reason, the principal investigator chose percentage agreement as the index of interrater reliability for all reliability analyses associated with items 7f – 7t.

Item 7u asked the question “Did the author(s) include a discussion related to the ‘acceptability’ of reliability criterion/values reported?” Coders responded “yes” or “no.” For this item, similar to the case with items 4a – 4d, coders responded “no” over 90% of the time, resulting in a lack of between items variance. Consequently, the principal investigator chose percentage agreement as the index of interrater reliability.

Items 7v1 to 7v12 were similar to the multiple response scenario described with item 1 and 7b. Items 7v1 to 7v6 asked coders to identify the lowest accepted reliability criterion reported and record the both the method used to calculate the criterion and the reliability criterion. Coders could also choose the option “standardized reliability not reported.” Similarly, items 7v7 to 7v12 asked coders to identify the highest accepted reliability criterion and follow the same procedure. For variables 7v1 (low accepted reliability method) and 7v7 (high accepted reliability method), all cases were included in the reliability analyses, as the option “standardized reliability not reported” was accounted for in these variables. For the subsequent items, only cases with more than one method were included in the reliability analyses. For the items related to the actual reliability values (7v2, 7v4, 7v6, 7v8, 7v10, and 7v12), only cases with reported reliability values were included in the reliability analyses.
For item 7v1, method used to calculate the lowest accepted reliability criterion, coders chose the option “standardized reliability not reported” over 60% of the time. For the item 7v7, method used to calculate the highest reliability criterion, coders chose the option “standardized reliability not reported over 70% of the time. This resulted in a lack of between items variance, which made percentage agreement the index of interrater reliability of choice for all reliability analyses associated with these items. The items relating to the second and third methods (items 7v3, 7v5, 7v9, and 7v11) of calculating reliability had small sample sizes (ranging from $n = 1$ to $n = 4$), which made percentage agreement the index of interrater reliability of choice for these items. Similarly, the items related to the actual reliability values (7v2, 7v4, 7v6, 7v8, 7v10, and 7v12) also had small sample sizes (ranging from $n = 1$ to $n = 15$) which also made percentage agreement the index of interrater reliability of choice for these items.

Item 7w asked the question “Was the information needed to code variable 7v clearly stated in the text?” Coders responded with one of three choices: “yes,” “no,” “standardized reliability not reported.” As in the case with items that were similar to this in nature, the principal investigator chose the generalizability coefficient for a design 2 generalizability study with an absolute (criterion-referenced) decision (ICC[A,k]) as the index of interrater reliability.

Item 7x asked coders to note the lowest reliability criterion deemed “acceptable” by the author. Coders noted the lowest acceptable reliability criterion (which ranged from 0.00 – 1.00) as an open-ended response. Coders could also choose the option “standardized reliability not reported.” Because the structure of this variable had both
categorical and quantitative components, the principal investigator chose percentage agreement as the index of interrater reliability.

“Acceptable” Reliability

As discussed in Chapter Two, there are no established standards for determining what signifies an acceptable level of reliability. In this study, the principal investigator used Neuendorf’s (2002) conventional rule that reliability coefficients of .80 or greater are acceptable in “most situations” (p.143) as a baseline. However, it is difficult to use a conventional rule that applies to reliability coefficients in general, when, as discussed in Chapter 2, each reliability coefficient has a separate mathematical basis and purpose. Further, generalizability theory coefficients for absolute decisions are often lower than coefficients for relative decisions because the error term accounts for rater variation in absolute decisions (Johanson, 2005).

Data Analysis

The principal investigator analyzed the data using SPSS software. In addition to the reliability analysis described in this chapter, the principal investigator used descriptive statistics to address the overarching research question and the 48 sub-problems that followed.
CHAPTER FOUR

Data Analysis

This chapter contains three main sections devoted to data analysis. The first section is a brief overview of the study. The second section presents interrater reliability analyses for all study variables. The third section presents an analysis of the sub-problems related to the overarching research question: How is intercoder reliability assessed and reported in published meta-analyses related to preschool through grade 12 education?

Overview

This study is a content analysis of meta-analyses published related to preschool through grade 12 education. The initial data set for this study included 100 journal articles acquired from the three databases: ERIC, Education Abstracts, and PsychInfo. During the coding process, a coder discovered that one article failed to meet inclusion criteria. This resulted in the exclusion of the article from the study. Therefore, the overall data set included 99 journal articles. Within these journal articles, 10 articles reported results from more than one meta-analysis. Because some authors reported interrater reliability information for more than one meta-analysis, each meta-analysis (whether the author reported interrater reliability information or not) was considered a separate case. Therefore, the final data set is comprised of 118 meta-analyses reported in 99 journal articles.
Interrater Reliability Analyses

This study used an item-by-item approach for all four reliability analyses described in Chapter Three: (a) initial reliability analysis; (b) final sample, first half reliability analysis; (c) final sample, second half reliability analysis; and (d) overall reliability analysis. Discussion in Chapter Three also describes the decision-making process for choosing the appropriate index of reliability for all variables in the study. Table 7 presents interrater reliability coefficients for items 1a – 1d and item 2. Table 8 presents interrater reliability coefficients for items 3a – 3i. Table 9 presents interrater reliability coefficients for items 4a – 4d. Table 10 presents interrater reliability coefficients for items 5a – 5d and items 6a – 6f. Table 11 presents interrater reliability coefficients for items 7a – 7x.

For the initial reliability sample, reliability coefficients based on ICC(A,k) ranged from 0.79 to 1.00 (\(M = 0.91, \ SD = 0.06\)), coefficients based on ICC(C,k) ranged from 0.86 to 1.00 (\(M = 0.95, \ SD = 0.06\)), and coefficients based on percentage agreement ranged from 0.00 to 1.00 (\(M = 0.63, \ SD = 0.36\)). For the final sample, first half reliability sample, reliability coefficients based on ICC(A,k) ranged from 0.87 to 1.00 (\(M = 0.98, \ SD = 0.04\)), coefficients based on ICC(C,k) ranged from 0.91 to 0.99 (\(M = 0.96, \ SD = 0.04\)), and coefficients based on percentage agreement ranged from 0.00 to 1.00 (\(M = 0.78, \ SD = 0.30\)). For the final sample, second half reliability sample, reliability coefficients based on ICC(A,k) ranged from 0.64 to 1.00 (\(M = 0.95, \ SD = 0.11\)), coefficients based on ICC(C,k) ranged from 0.34 to 1.00 (\(M = 0.78, \ SD = 0.34\)), and coefficients based on percentage agreement ranged from 0.00 to 1.00 (\(M = 0.61, \ SD = 0.35\)). For the overall reliability sample, reliability coefficients based on ICC(A,k) ranged
from 0.85 to 0.99 ($M = 0.93$, $SD = 0.45$), coefficients based on ICC(C,k) ranged from 0.87 to 1.00 ($M = 0.95$, $SD = 0.06$), and coefficients based on percentage agreement ranged from 0.00 to 1.00 ($M = 0.66$, $SD = 0.34$).
Table 7

Interrater Reliability for Items 1a – 1d and Item 2

<table>
<thead>
<tr>
<th>Item</th>
<th>Final Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Initial (n)</td>
</tr>
<tr>
<td>1a. Content area: choice 1</td>
<td>1.00&lt;sup&gt;a&lt;/sup&gt; (25)</td>
</tr>
<tr>
<td>1b. Content area: choice 2</td>
<td>0.58&lt;sup&gt;c&lt;/sup&gt; (24)</td>
</tr>
<tr>
<td>1c. Content area: choice 3</td>
<td>0.00&lt;sup&gt;c&lt;/sup&gt; (16)</td>
</tr>
<tr>
<td>1d. Content area: choice 4</td>
<td>0.00&lt;sup&gt;c&lt;/sup&gt; (2)</td>
</tr>
<tr>
<td>2. Number of primary studies</td>
<td>1.00&lt;sup&gt;b&lt;/sup&gt; (25)</td>
</tr>
</tbody>
</table>

Note. Dashes indicate that interrater reliability was not estimated because there were no cases.

<sup>a</sup>ICC(A,k). <sup>b</sup>ICC(C,k). <sup>c</sup>Percent Agreement expressed as a proportion.
Table 8

*Interrater Reliability for Items 3a – 3i*

<table>
<thead>
<tr>
<th>Item</th>
<th>Initial (n)</th>
<th>Half One (n)</th>
<th>Half Two (n)</th>
<th>Overall (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3a. Report number of coders</td>
<td>0.92&lt;sup&gt;a&lt;/sup&gt; (25)</td>
<td>1.00&lt;sup&gt;a&lt;/sup&gt; (9)</td>
<td>1.00&lt;sup&gt;a&lt;/sup&gt; (8)</td>
<td>0.95&lt;sup&gt;a&lt;/sup&gt; (42)</td>
</tr>
<tr>
<td>3b1. Number of coders – general</td>
<td>0.96&lt;sup&gt;a&lt;/sup&gt; (25)</td>
<td>1.00&lt;sup&gt;a&lt;/sup&gt; (9)</td>
<td>1.00&lt;sup&gt;a&lt;/sup&gt; (8)</td>
<td>0.98&lt;sup&gt;a&lt;/sup&gt; (42)</td>
</tr>
<tr>
<td>3b2. Number of coders – specific</td>
<td>0.75&lt;sup&gt;c&lt;/sup&gt; (12)</td>
<td>1.00&lt;sup&gt;c&lt;/sup&gt; (4)</td>
<td>0.67&lt;sup&gt;c&lt;/sup&gt; (2)</td>
<td>0.78&lt;sup&gt;c&lt;/sup&gt; (18)</td>
</tr>
<tr>
<td>3c. Clarity: reporting number of coders</td>
<td>0.88&lt;sup&gt;a&lt;/sup&gt; (25)</td>
<td>1.00&lt;sup&gt;a&lt;/sup&gt; (9)</td>
<td>1.00&lt;sup&gt;a&lt;/sup&gt; (8)</td>
<td>0.94&lt;sup&gt;a&lt;/sup&gt; (42)</td>
</tr>
<tr>
<td>3d. Differentiation: reliability vs. actual</td>
<td>0.82&lt;sup&gt;a&lt;/sup&gt; (25)</td>
<td>1.00&lt;sup&gt;a&lt;/sup&gt; (9)</td>
<td>1.00&lt;sup&gt;a&lt;/sup&gt; (8)</td>
<td>0.91&lt;sup&gt;a&lt;/sup&gt; (42)</td>
</tr>
<tr>
<td>3e. Size of reliability sample</td>
<td>0.91&lt;sup&gt;a&lt;/sup&gt; (8)</td>
<td>1.00&lt;sup&gt;a&lt;/sup&gt; (4)</td>
<td>0.67&lt;sup&gt;c&lt;/sup&gt; (3)</td>
<td>0.97&lt;sup&gt;a&lt;/sup&gt; (15)</td>
</tr>
<tr>
<td>3f. Clarity: reporting reliability sample size</td>
<td>0.92&lt;sup&gt;a&lt;/sup&gt; (25)</td>
<td>0.97&lt;sup&gt;a&lt;/sup&gt; (9)</td>
<td>0.92&lt;sup&gt;a&lt;/sup&gt; (8)</td>
<td>0.93&lt;sup&gt;a&lt;/sup&gt; (42)</td>
</tr>
<tr>
<td>3g. Reliability sample: proportion of studies</td>
<td>0.99&lt;sup&gt;b&lt;/sup&gt; (8)</td>
<td>0.99&lt;sup&gt;b&lt;/sup&gt; (4)</td>
<td>1.00&lt;sup&gt;c&lt;/sup&gt; (3)</td>
<td>0.99&lt;sup&gt;b&lt;/sup&gt; (15)</td>
</tr>
<tr>
<td>3h. Report description of coder expertise</td>
<td>0.79&lt;sup&gt;a&lt;/sup&gt; (25)</td>
<td>1.00&lt;sup&gt;a&lt;/sup&gt; (9)</td>
<td>1.00&lt;sup&gt;a&lt;/sup&gt; (8)</td>
<td>0.85&lt;sup&gt;a&lt;/sup&gt; (42)</td>
</tr>
<tr>
<td>3i1. Coder’s levels of expertise: choice 1</td>
<td>0.72&lt;sup&gt;c&lt;/sup&gt; (25)</td>
<td>1.00&lt;sup&gt;c&lt;/sup&gt; (9)</td>
<td>1.00&lt;sup&gt;c&lt;/sup&gt; (8)</td>
<td>0.83&lt;sup&gt;c&lt;/sup&gt; (42)</td>
</tr>
<tr>
<td>3i2. Coder’s levels of expertise: choice 2</td>
<td>0.84&lt;sup&gt;c&lt;/sup&gt; (25)</td>
<td>1.00&lt;sup&gt;c&lt;/sup&gt; (9)</td>
<td>1.00&lt;sup&gt;c&lt;/sup&gt; (8)</td>
<td>0.90&lt;sup&gt;c&lt;/sup&gt; (42)</td>
</tr>
</tbody>
</table>

<sup>a</sup>ICC(A,k). <sup>b</sup>ICC(C,k). <sup>c</sup>Percent Agreement expressed as a proportion.
Table 9

*Interrater Reliability for Items 4a – 4d*

<table>
<thead>
<tr>
<th>Item</th>
<th>Initial (n)</th>
<th>Half One (n)</th>
<th>Half Two (n)</th>
<th>Overall (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4a. Report coder training</td>
<td>0.88 (25)</td>
<td>1.00 (9)</td>
<td>1.00 (8)</td>
<td>0.93 (42)</td>
</tr>
<tr>
<td>4b1. Amount of coder training: choice 1</td>
<td>0.88 (25)</td>
<td>1.00 (9)</td>
<td>1.00 (8)</td>
<td>0.93 (42)</td>
</tr>
<tr>
<td>4b2. Units of coder training: choice 1</td>
<td>0.25 (4)</td>
<td>–</td>
<td>–</td>
<td>0.25 (4)</td>
</tr>
<tr>
<td>4b3. Amount of coder training: choice 2</td>
<td>0.00 (1)</td>
<td>–</td>
<td>–</td>
<td>0.00 (1)</td>
</tr>
<tr>
<td>4b4. Units of coder training: choice 2</td>
<td>0.00 (1)</td>
<td>–</td>
<td>–</td>
<td>0.00 (1)</td>
</tr>
<tr>
<td>4c. Description of coder training</td>
<td>0.84 (25)</td>
<td>1.00 (9)</td>
<td>1.00 (8)</td>
<td>0.90 (42)</td>
</tr>
<tr>
<td>4d1. Report stopping rule</td>
<td>0.96 (25)</td>
<td>0.89 (9)</td>
<td>1.00 (8)</td>
<td>0.95 (42)</td>
</tr>
<tr>
<td>4d2. Stopping rule criterion</td>
<td>0.00 (1)</td>
<td>0.00 (1)</td>
<td>–</td>
<td>0.00 (2)</td>
</tr>
</tbody>
</table>

*Note.* All reliability coefficients estimated by percent agreement. Values shown are percentage agreement expressed as a proportion. Dashes indicate that interrater reliability was not estimated because there were no cases.
Table 10

*Interrater Reliability for Items 5a – 5d and 6a – 6f*

<table>
<thead>
<tr>
<th>Item</th>
<th>Initial (n)</th>
<th>Half One (n)</th>
<th>Half Two (n)</th>
<th>Overall (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5a. Description of code book</td>
<td>0.95a (25)</td>
<td>0.87a (9)</td>
<td>0.64a (8)</td>
<td>0.89a (42)</td>
</tr>
<tr>
<td>5b. Report code book available</td>
<td>1.00c (25)</td>
<td>1.00c (9)</td>
<td>0.88c (8)</td>
<td>0.98c (42)</td>
</tr>
<tr>
<td>5c. Report resolution: coding discrepancies</td>
<td>0.96c (25)</td>
<td>1.00c (9)</td>
<td>0.88c (8)</td>
<td>0.95c (42)</td>
</tr>
<tr>
<td>5d. Description of resolution method</td>
<td>0.93a (25)</td>
<td>0.96a (9)</td>
<td>0.75c (8)</td>
<td>0.88c (42)</td>
</tr>
<tr>
<td>6a. Article addresses reliability</td>
<td>0.95a (25)</td>
<td>1.00a (9)</td>
<td>1.00a (8)</td>
<td>0.97a (42)</td>
</tr>
<tr>
<td>6b. Presents reliability information in text</td>
<td>0.90c (25)</td>
<td>0.88c (9)</td>
<td>0.89c (8)</td>
<td>1.00c (42)</td>
</tr>
<tr>
<td>6c. Number of sentences: discussion</td>
<td>0.86b (25)</td>
<td>0.96b (9)</td>
<td>0.34b (8)</td>
<td>0.87b (42)</td>
</tr>
<tr>
<td>6d. Number of sentences: reporting</td>
<td>0.95b (25)</td>
<td>0.91b (9)</td>
<td>1.00b (8)</td>
<td>0.93b (42)</td>
</tr>
<tr>
<td>6e. Include citation(s) about reliability</td>
<td>1.00c (25)</td>
<td>1.00c (9)</td>
<td>0.88c (8)</td>
<td>0.98c (42)</td>
</tr>
<tr>
<td>6f. Number of citation(s) reported</td>
<td>1.00c (25)</td>
<td>1.00c (9)</td>
<td>0.88c (8)</td>
<td>0.98c (42)</td>
</tr>
</tbody>
</table>

*Note.* The term “reliability” indicates interrater reliability.

*aICCA(k).*  
*bICCC(k).*  
*cPercent Agreement expressed as a proportion.*
Table 11

*Interrater Reliability for Items 7a – 7x*

<table>
<thead>
<tr>
<th>Item</th>
<th>Final Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item</td>
<td>Initial (n)</td>
</tr>
<tr>
<td>7a. Report specific reliability method</td>
<td>0.92^c (25)</td>
</tr>
<tr>
<td>7b1. Specific reliability method: choice 1</td>
<td>0.87^a (25)</td>
</tr>
<tr>
<td>7b2. Specific reliability method: choice 2</td>
<td>0.33^c (3)</td>
</tr>
<tr>
<td>7b3. Specific reliability method: choice 3</td>
<td>0.00^c (1)</td>
</tr>
<tr>
<td>7c. Method of reporting reliability statistics</td>
<td>0.90^a (25)</td>
</tr>
<tr>
<td>7e. Differentiate effect size vs. study descriptor coding</td>
<td>0.96^c (25)</td>
</tr>
<tr>
<td>7f. Effect size: Overall/average reliability</td>
<td>0.50^c (2)</td>
</tr>
<tr>
<td>7g. Effect size: Range of reliability</td>
<td>0.00^c (2)</td>
</tr>
<tr>
<td>7h. Effect size: Specific reliability</td>
<td>0.50^c (2)</td>
</tr>
<tr>
<td>7i. Effect size: Reliability for all variables</td>
<td>0.00^c (2)</td>
</tr>
</tbody>
</table>

*(table continues)*
Table 11 (continued)

**Interrater Reliability for Items 7a – 7x**

<table>
<thead>
<tr>
<th>Item</th>
<th>Initial (n)</th>
<th>Half One (n)</th>
<th>Half Two (n)</th>
<th>Overall (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>7j. Effect size: Clarity of reporting</td>
<td>0.00&lt;sup&gt;c&lt;/sup&gt; (2)</td>
<td>0.67&lt;sup&gt;c&lt;/sup&gt; (3)</td>
<td>0.25&lt;sup&gt;c&lt;/sup&gt; (4)</td>
<td>0.33&lt;sup&gt;c&lt;/sup&gt; (9)</td>
</tr>
<tr>
<td>7k. Descriptor: Overall/average reliability</td>
<td>0.50&lt;sup&gt;c&lt;/sup&gt; (2)</td>
<td>0.33&lt;sup&gt;c&lt;/sup&gt; (3)</td>
<td>0.25&lt;sup&gt;c&lt;/sup&gt; (4)</td>
<td>0.33&lt;sup&gt;c&lt;/sup&gt; (9)</td>
</tr>
<tr>
<td>7l. Descriptor: Range of reliability</td>
<td>0.50&lt;sup&gt;c&lt;/sup&gt; (2)</td>
<td>0.67&lt;sup&gt;c&lt;/sup&gt; (3)</td>
<td>0.25&lt;sup&gt;c&lt;/sup&gt; (4)</td>
<td>0.44&lt;sup&gt;c&lt;/sup&gt; (9)</td>
</tr>
<tr>
<td>7m. Descriptor: Specific reliability</td>
<td>0.50&lt;sup&gt;c&lt;/sup&gt; (2)</td>
<td>0.00&lt;sup&gt;c&lt;/sup&gt; (3)</td>
<td>0.00&lt;sup&gt;c&lt;/sup&gt; (4)</td>
<td>0.11&lt;sup&gt;c&lt;/sup&gt; (9)</td>
</tr>
<tr>
<td>7n. Descriptor: Reliability for all variables</td>
<td>0.00&lt;sup&gt;c&lt;/sup&gt; (2)</td>
<td>0.33&lt;sup&gt;c&lt;/sup&gt; (3)</td>
<td>0.25&lt;sup&gt;c&lt;/sup&gt; (4)</td>
<td>0.22&lt;sup&gt;c&lt;/sup&gt; (9)</td>
</tr>
<tr>
<td>7o. Descriptor: Clarity of reporting</td>
<td>0.50&lt;sup&gt;c&lt;/sup&gt; (2)</td>
<td>0.67&lt;sup&gt;c&lt;/sup&gt; (3)</td>
<td>0.25&lt;sup&gt;c&lt;/sup&gt; (4)</td>
<td>0.44&lt;sup&gt;c&lt;/sup&gt; (9)</td>
</tr>
<tr>
<td>7p. General: Overall/average reliability</td>
<td>0.88&lt;sup&gt;c&lt;/sup&gt; (24)</td>
<td>0.86&lt;sup&gt;c&lt;/sup&gt; (7)</td>
<td>0.57&lt;sup&gt;c&lt;/sup&gt; (7)</td>
<td>0.82&lt;sup&gt;c&lt;/sup&gt; (38)</td>
</tr>
<tr>
<td>7q. General: Range of reliability</td>
<td>0.92&lt;sup&gt;c&lt;/sup&gt; (24)</td>
<td>0.86&lt;sup&gt;c&lt;/sup&gt; (7)</td>
<td>0.57&lt;sup&gt;c&lt;/sup&gt; (7)</td>
<td>0.84&lt;sup&gt;c&lt;/sup&gt; (38)</td>
</tr>
<tr>
<td>7r. General: Specific reliability</td>
<td>0.88&lt;sup&gt;c&lt;/sup&gt; (24)</td>
<td>0.71&lt;sup&gt;c&lt;/sup&gt; (7)</td>
<td>0.43&lt;sup&gt;c&lt;/sup&gt; (7)</td>
<td>0.76&lt;sup&gt;c&lt;/sup&gt; (38)</td>
</tr>
<tr>
<td>7s. General: Reliability for all variables</td>
<td>0.83&lt;sup&gt;c&lt;/sup&gt; (24)</td>
<td>0.86&lt;sup&gt;c&lt;/sup&gt; (7)</td>
<td>0.57&lt;sup&gt;c&lt;/sup&gt; (7)</td>
<td>0.79&lt;sup&gt;c&lt;/sup&gt; (38)</td>
</tr>
<tr>
<td>7t. General: Clarity of reporting</td>
<td>0.79&lt;sup&gt;c&lt;/sup&gt; (24)</td>
<td>0.86&lt;sup&gt;c&lt;/sup&gt; (7)</td>
<td>0.43&lt;sup&gt;c&lt;/sup&gt; (7)</td>
<td>0.74&lt;sup&gt;c&lt;/sup&gt; (38)</td>
</tr>
</tbody>
</table>

*(table continues)*
Table 11 (continued)

*Interrater Reliability for Items 7a – 7x*

<table>
<thead>
<tr>
<th>Item</th>
<th>Final Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Initial ((n))</td>
</tr>
<tr>
<td>7u. “Acceptability” of reliability values</td>
<td>1.00(^c) (25)</td>
</tr>
<tr>
<td>7v1. Low accepted method: choice 1</td>
<td>0.84(^c) (25)</td>
</tr>
<tr>
<td>7v2. Low accepted value: choice 1</td>
<td>0.75(^c) (8)</td>
</tr>
<tr>
<td>7v3. Low accepted method: choice 2</td>
<td>0.67(^c) (3)</td>
</tr>
<tr>
<td>7v4. Low accepted value: choice 2</td>
<td>0.67(^c) (3)</td>
</tr>
<tr>
<td>7v5. Low accepted method: choice 3</td>
<td>1.00(^c) (1)</td>
</tr>
<tr>
<td>7v6. Low accepted value: choice 3</td>
<td>1.00(^c) (1)</td>
</tr>
<tr>
<td>7v7. High accepted method: choice 1</td>
<td>0.84(^c) (25)</td>
</tr>
<tr>
<td>7v8. High accepted value: choice 1</td>
<td>0.75(^c) (8)</td>
</tr>
<tr>
<td>7v9. High accepted method: choice 2</td>
<td>0.67(^c) (3)</td>
</tr>
<tr>
<td>7v10. High accepted value: choice 2</td>
<td>0.67(^c) (3)</td>
</tr>
</tbody>
</table>

*(table continues)*
Table 11 (continued)

*Interrater Reliability for Items 7a – 7x*

<table>
<thead>
<tr>
<th>Item</th>
<th>Initial (n)</th>
<th>Half One (n)</th>
<th>Half Two (n)</th>
<th>Overall (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>7v11. High accepted method: choice 3</td>
<td>1.00(^c) (1)</td>
<td>–</td>
<td>–</td>
<td>1.00(^c) (1)</td>
</tr>
<tr>
<td>7v12. High accepted value: choice 3</td>
<td>1.00(^c) (1)</td>
<td>–</td>
<td>–</td>
<td>1.00(^c) (1)</td>
</tr>
<tr>
<td>7w. Clarity of reporting (accepted reliability values)</td>
<td>0.90(^a) (25)</td>
<td>1.00(^a) (9)</td>
<td>0.88(^c) (8)</td>
<td>0.85(^a) (42)</td>
</tr>
<tr>
<td>7x. Lowest accepted reliability value</td>
<td>0.96(^c) (25)</td>
<td>1.00(^c) (9)</td>
<td>1.00(^c) (8)</td>
<td>0.98(^c) (42)</td>
</tr>
</tbody>
</table>

*Note.* The term “reliability” indicates interrater reliability. Dashes indicate that interrater reliability was not estimated because there were no cases.

\(^a\)ICC(A,k). \(^b\)ICC(C,k). \(^c\)Percent Agreement expressed as a proportion. \(^d\)Variable 7d is a string variable, therefore interrater reliability was not calculated.
Analysis of the Sub-problems

Data analyses that pertain to the overarching research question: How is intercoder reliability assessed and reported in published meta-analyses related to preschool through grade 12 education? are presented in this section. This study examined 48 sub-problems in order to provide a detailed response to the research question. The sub-problems do not represent a set of criteria for judging the quality of intercoder reliability reporting in meta-analyses. They simply serve as a tool to explore the current state of intercoder reliability reporting among published meta-analyses related to preschool through grade 12 education. Detailed analyses for each sub-problem follow.

What content areas in education are utilizing meta-analysis? Table 12 contains data regarding the content areas represented by the meta-analyses included in the study. Because this item was a multiple response item, coders could choose more than one content area per meta-analysis. Therefore, Table 12 describes the responses in two ways: responses by content area (N = 172) and responses as a percent of total cases (N = 118). For the responses by category, coders identified 172 content areas for the 118 meta-analyses in the study. Therefore, this column represents a breakdown of the percent of responses by content area using the total number of responses as the denominator. The percent of total cases column represents the percent of meta-analyses assigned to a given content area. Because this column uses the total number of meta-analyses as the denominator, the column does not sum to 100%.

For 66 of the meta-analyses in the study (55.9%), coders identified a single content area. Coders identified two content areas for 50 meta-analyses (42.4%) and three content areas for 2 meta-analyses (1.7%) in the study. As demonstrated by data in Table
12, language arts education was the content area that had the most published meta-analyses, with 25.4% of the meta-analyses in the study related to this content area. Special education and academic (educational) achievement were the second most represented content areas, with 24.6% of the meta-analyses in the study related to each content area. Arts education was the third most represented content area (18.6%) followed by the category labeled “other” (16.1%). The category labeled “other” included topics such as dynamic assessment, childhood sexual abuse prevention programs, and school mobility. The remaining content areas each accounted for less than 6% of the meta-analyses in the study.
<table>
<thead>
<tr>
<th>Content Area</th>
<th>Responses by Category</th>
<th>Percent</th>
<th>Total Cases^b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academic Leadership</td>
<td>3</td>
<td>1.7</td>
<td>2.5</td>
</tr>
<tr>
<td>Academic (Educational) Achievement</td>
<td>29</td>
<td>16.9</td>
<td>24.6</td>
</tr>
<tr>
<td>Arts Education</td>
<td>22</td>
<td>12.8</td>
<td>18.6</td>
</tr>
<tr>
<td>Computer-Assisted Instruction</td>
<td>7</td>
<td>4.1</td>
<td>5.9</td>
</tr>
<tr>
<td>English as a Second Language</td>
<td>1</td>
<td>0.6</td>
<td>0.8</td>
</tr>
<tr>
<td>Foreign Language Education</td>
<td>2</td>
<td>1.2</td>
<td>1.7</td>
</tr>
<tr>
<td>General Early Childhood Education</td>
<td>4</td>
<td>2.3</td>
<td>3.4</td>
</tr>
<tr>
<td>General Elementary Education</td>
<td>2</td>
<td>1.2</td>
<td>1.7</td>
</tr>
<tr>
<td>Health Education</td>
<td>3</td>
<td>1.7</td>
<td>2.5</td>
</tr>
<tr>
<td>Language Arts Education</td>
<td>30</td>
<td>17.4</td>
<td>25.4</td>
</tr>
<tr>
<td>Mathematics Education</td>
<td>7</td>
<td>4.1</td>
<td>5.9</td>
</tr>
<tr>
<td>Physical Education</td>
<td>1</td>
<td>0.6</td>
<td>0.8</td>
</tr>
<tr>
<td>Science Education</td>
<td>6</td>
<td>3.5</td>
<td>5.1</td>
</tr>
<tr>
<td>Special Education</td>
<td>29</td>
<td>16.9</td>
<td>24.6</td>
</tr>
<tr>
<td>School Psychology/Guidance Counseling</td>
<td>3</td>
<td>1.7</td>
<td>2.5</td>
</tr>
<tr>
<td>Technical Education</td>
<td>1</td>
<td>0.6</td>
<td>0.8</td>
</tr>
<tr>
<td>Other</td>
<td>19</td>
<td>11.0</td>
<td>16.1</td>
</tr>
<tr>
<td>Violence Prevention</td>
<td>3</td>
<td>1.7</td>
<td>2.5</td>
</tr>
</tbody>
</table>

^aN = 172. ^bColumn does not sum to 100% due to multiple responses (N = 118).
What is the average number of studies included in a meta-analysis related to education? Authors for 115 out of the 118 meta-analyses in the study reported the number of studies included in the meta-analysis. The number of studies included in a meta-analysis related to education ranged from 3 to 267 studies. The mean number of studies included in a meta-analysis related to education was 43.1 (SD = 50.3). The median number of studies included in a meta-analysis related to education was 25.0. Table 13 presents a group frequency distribution of the number of studies included in a meta-analysis related to education.

Table 13

<table>
<thead>
<tr>
<th>Number of Studies Included in Meta-Analyses Related to Education</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Studies (n= 115)</td>
</tr>
<tr>
<td>---------------------------</td>
</tr>
<tr>
<td>Less than 10</td>
</tr>
<tr>
<td>10 – 19</td>
</tr>
<tr>
<td>20 – 29</td>
</tr>
<tr>
<td>30 – 39</td>
</tr>
<tr>
<td>40 – 49</td>
</tr>
<tr>
<td>50 – 59</td>
</tr>
<tr>
<td>60 – 99</td>
</tr>
<tr>
<td>100 or More</td>
</tr>
</tbody>
</table>
Did the authors report the number of coders that coded each study? Of the 118 meta-analyses in the study, 50 meta-analyses (42.4%) reported the number of coders that coded each study.

How many coders rated each study? Fifty of the 118 meta-analyses (42.4%) in the study reported the number of coders that coded each study. Of these meta-analyses, 4.0% reported using one coder and 96.0% of the meta-analyses reported using more than one coder. Table 14 presents a frequency table detailing the specific number of coders reported by authors reporting more than one coder.

Of the 68 meta-analyses that did not report the number of coders that coded each study (57.6% of the total sample), it could be inferred that more than one coder was utilized in 13 (19.1%) of the studies. This inference was possible because meta-analyses referred to “interrater reliability,” which implies more than one coder coded the studies in the meta-analysis. In the remaining 55 studies, meta-analyses did not report any details about the person(s) involved in the coding process.

Table 14

Specific Number of Coders

<table>
<thead>
<tr>
<th>Number of Studies (n = 50)</th>
<th>Percent (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.0% (2)</td>
</tr>
<tr>
<td>2</td>
<td>66.0% (33)</td>
</tr>
<tr>
<td>3</td>
<td>18.0% (9)</td>
</tr>
<tr>
<td>4</td>
<td>10.0% (5)</td>
</tr>
<tr>
<td>5</td>
<td>2.0% (1)</td>
</tr>
</tbody>
</table>
How clear (easy to find) was the information presented regarding the number of coders? Coders responded to this item for all 118 meta-analyses in the study. In 42.4% of the meta-analyses, coders reported that “Yes, the information was clearly stated (easy to find).” In 11.9% of the meta-analyses, coders reported that “No, the information was not clear (easy to find). I had to make assumptions to code this variable.” In the remaining 45.8% of the meta-analyses, coders reported, “I was unable to ascertain any information about the individuals involved in coding the primary studies.”

What percentage of authors differentiate between the coding of the reliability sample and the coding of the actual sample? Of the 118 meta-analyses in the study, 47 meta-analyses (39.8%) differentiated between the coding of the reliability sample and the coding for the actual sample.

What is the average size of the reliability sample? Of the 47 meta-analyses (39.8% of the total sample) differentiating between the coding of the reliability sample and the coding of the actual sample, the mean size of the reliability sample was 34.9 ($SD = 43.9$). The median size of the reliability sample was 19.0. The size of the reliability sample ranged from 2 to 232 studies. Table 15 presents a group frequency distribution of the number of studies included in the reliability sample.
Table 15

<table>
<thead>
<tr>
<th>Number of Studies Included in the Reliability Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Studies (n = 47)</td>
</tr>
<tr>
<td>Less than 10</td>
</tr>
<tr>
<td>10 – 19</td>
</tr>
<tr>
<td>20 – 29</td>
</tr>
<tr>
<td>30 – 39</td>
</tr>
<tr>
<td>40 – 49</td>
</tr>
<tr>
<td>50 or More</td>
</tr>
</tbody>
</table>

How clear (easy to find) was the information presented regarding the size of the reliability sample? Coders responded to this item for all 118 meta-analyses in the study. In 33.1% of the meta-analyses, coders reported that “Yes, the information was clearly stated (easy to find).” In 6.8% of the meta-analyses, coders reported that “No, the information was not clear (easy to find). I had to make assumptions to code this variable.” In the remaining 60.2% of the meta-analyses, coders reported, “The authors did not report any information regarding the assessment of reliability.”

What proportion of primary studies was included in the reliability sample? Of meta-analyses reporting the number of primary studies included in the reliability sample and the total number of studies in the meta-analysis (n = 47, or 39.8% of the total sample), the mean proportion of primary studies included in the reliability sample was 0.75 (SD = 0.37). The proportion of studies included in the reliability sample ranged from 0.05 to 1.00, with 32 of the 47 studies (68.1%) using all primary studies in the sample in
the reliability analysis. Of these 32 studies, the number of primary studies ranged from 7 to 225 studies, with a mean of 45.4 ($SD = 49.7$) and a median of 22.0. Table 16 presents a group frequency distribution of the proportion of studies included in the reliability sample.

Table 16

<table>
<thead>
<tr>
<th>Proportion of Studies (n= 47)</th>
<th>Percent (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than 0.10</td>
<td>4.3% (2)</td>
</tr>
<tr>
<td>0.10 – 0.19</td>
<td>8.5% (4)</td>
</tr>
<tr>
<td>0.20 – 0.29</td>
<td>12.8% (6)</td>
</tr>
<tr>
<td>0.30 – 0.39</td>
<td>6.4% (3)</td>
</tr>
<tr>
<td>1.0</td>
<td>68.1% (32)</td>
</tr>
</tbody>
</table>

*Did the authors provide a description of the coders’ levels of expertise?* Of the 118 meta-analyses in the study, 24 meta-analyses (20.3%) provided a description of the coders’ levels of expertise.

*Who coded the studies (e.g., undergraduate research assistants, graduate research assistants, study authors, etc.)?* Of the meta-analyses reporting a description of the coders’ levels of expertise ($n = 24$, or 20.3% of the total sample), the coders were described as “study authors” in 62.5% of the meta-analyses, with 11 of the 24 meta-analyses (45.8%) exclusively describing the coders as such. Meta-analyses reported graduate students or graduate research assistants as coders in 33.3% of the meta-analyses,
with 7 of the 24 meta-analyses (29.2%) exclusively describing the coders as such. Meta-analyses used the terms “trained research assistant,” “research assistant,” or “researcher” to describe coders in 20.8% of the meta-analyses. Note that the percentages do not sum to 100% because some authors described more than one coder per meta-analysis. For example, 2 of the 24 meta-analyses (8.3%), described coders as both “study authors” and “trained research assistants.” Table 17 presents a frequency distribution for the description of coders.

Table 17

<table>
<thead>
<tr>
<th>Coder Description</th>
<th>Responses by Category</th>
<th>Percent of Total Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graduate Student/Graduate Research Assistant</td>
<td>8</td>
<td>28.6</td>
</tr>
<tr>
<td>Research Assistant</td>
<td>1</td>
<td>3.6</td>
</tr>
<tr>
<td>Researcher</td>
<td>1</td>
<td>3.6</td>
</tr>
<tr>
<td>Study Author</td>
<td>15</td>
<td>53.6</td>
</tr>
<tr>
<td>Trained Research Assistant</td>
<td>3</td>
<td>10.7</td>
</tr>
</tbody>
</table>

^aColumn does not sum to 100% due to multiple responses (n = 24).

Did the authors report coder training procedures? Of the 118 meta-analyses in the study, 2 meta-analyses (1.7%) reported coder training procedures.
How much training do coders receive? Two meta-analyses reported coder training procedures. One meta-analysis described a one-session training lasting two hours in length. The other meta-analysis described a one-session training lasting three hours in length.

How do authors describe coder training procedures? This item used a 5-point Likert scale where “1” corresponded to “Coder training is kept secret; not described at all” and “5” represented “Coder training is described in enough detail that it would be easy to replicate the coder training procedures.” Of the 118 meta-analyses in the study, coder training procedures were described as “Coder training kept secret; not described at all” 96.6% of the time. Table 18 presents a frequency table with detailed responses for this item.

Table 18

Description of Coder Training Procedures

<table>
<thead>
<tr>
<th>Description of Studies (N = 118)</th>
<th>Percent (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coder training is kept secret; not described at all</td>
<td>96.6% (114)</td>
</tr>
<tr>
<td>The author acknowledges coder training, but provides no details</td>
<td>1.7% (2)</td>
</tr>
<tr>
<td>A brief description of coder training is provided</td>
<td>1.7% (2)</td>
</tr>
<tr>
<td>Coder training is described, but additional research is necessary to replicate coder training procedures</td>
<td>0.0% (0)</td>
</tr>
<tr>
<td>Coder training is described in enough detail that it would be easy to replicate coder training procedures</td>
<td>0.0% (0)</td>
</tr>
</tbody>
</table>
What is the criterion for consistency (e.g., stopping rule) to determine that coders were adequately trained? Of the 118 meta-analyses in the study, 3 meta-analyses (2.5%) reported using a criterion for consistency (e.g., stopping rule) to determined that coders were adequately trained. Of the three meta-analyses reporting a criterion for consistency, two meta-analyses used 0.8 as the stopping rule and one meta-analysis used 0.9 as the stopping rule.

How do authors describe the code book? This item used a 5-point Likert scale where “1” corresponded to “Code book is kept secret; not described at all” and “5” represented “Code book is described in enough detail that it would be easy to replicate the study.” In general, coders used the following conventions to code this variable. If the study did not mention the use of a code book, coders assigned “code book is kept secret; not described at all.” If the author mentioned a code book, but provided no details pertaining to specific variables, coders assigned “the author acknowledges use of a code book but provides no details.” If the author provided a brief text description of code book, but provided no details regarding specific variables, the coders assigned “brief description of the code book is provided.” Coders assigned the designation “code book is described but additional research is necessary to replicate the study” only if the study included a table or paragraph listing most variables, indicating that more research was necessary to define the variables. Finally, coders reserved the designation “code book is described in enough detail that it would be easy to replicate the study” only for those meta-analyses which provided a listing of variables included in the study and their operational definitions.
Of the 118 meta-analyses in the study, coders most frequently (44.9%) assessed authors’ descriptions of the code book as “Code book is described, but additional research is necessary to replicate the study,” followed by “A brief description of the code book is provided” (21.2%). Additionally, 30.5% of the time, coders assessed authors’ description of the code book as “Code book is kept secret, not described at all.” Eight of the 118 meta-analyses (6.8%) indicated that a copy of the code book (or coding protocol) was available upon request. Table 19 presents a frequency table with detailed responses for this item.
Table 19

*Description of the Code Book*

<table>
<thead>
<tr>
<th>Description of Studies (N = 118)</th>
<th>Percent (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Code book is kept secret; not described at all</td>
<td>30.5% (36)</td>
</tr>
<tr>
<td>The author acknowledges use of a code book, but provides no details</td>
<td>1.7% (2)</td>
</tr>
<tr>
<td>A brief description of the code book is provided</td>
<td>21.2% (25)</td>
</tr>
<tr>
<td>Code book is described, but additional research is necessary to replicate the study</td>
<td>44.9% (53)</td>
</tr>
<tr>
<td>Code book is described in enough detail that it would be easy to replicate the study</td>
<td>1.7% (2)</td>
</tr>
</tbody>
</table>

*Do authors report how coder discrepancies were resolved?* Of the 118 meta-analyses in the study, 36 meta-analyses (30.5%) reported how coder discrepancies were resolved.

*How do authors describe the procedures in place to resolve coder discrepancies?* This item used a 5-point Likert scale where “1” corresponded to “Method of resolution is kept secret; not described at all” and “5” represented “Method of resolution is described in enough detail that it would be easy to replicate.” In general, coders utilized conventions that paralleled that of the Likert scaled items pertaining to the code book to code this variable.
Of the 118 meta-analyses in the study, coders described the method to resolve coder discrepancies as “Method of resolution is kept secret, not described at all” 69.5% of the time. Table 20 presents a frequency table with detailed responses for this item.

Table 20

*Description of Method Used to Resolve Coding Discrepancies*

<table>
<thead>
<tr>
<th>Description of Studies ( (N = 118) )</th>
<th>Percent ( (n) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method of resolution is kept secret; not described at all</td>
<td>69.5% (82)</td>
</tr>
<tr>
<td>The author acknowledges discrepancies were resolved, but</td>
<td></td>
</tr>
<tr>
<td>provides no details</td>
<td>1.7% (2)</td>
</tr>
<tr>
<td>A brief description of the method of resolution is provided</td>
<td>28.0% (33)</td>
</tr>
<tr>
<td>Method of resolution is described, but additional research is</td>
<td></td>
</tr>
<tr>
<td>necessary to replicate the resolution method</td>
<td>0.8% (1)</td>
</tr>
<tr>
<td>Method of resolution is described in enough detail that it would</td>
<td></td>
</tr>
<tr>
<td>be easy to replicate the study</td>
<td>0.0% (0)</td>
</tr>
</tbody>
</table>

*What percentage of meta-analyses acknowledge intercoder reliability?* Of the 118 meta-analyses in the study, 47 meta-analyses (39.8%) acknowledged intercoder reliability by either generally discussing or by reporting intercoder reliability analyses. Further, 35 meta-analyses (29.7%) both generally discussed intercoder reliability and reported intercoder reliability analyses.

*How do authors present intercoder reliability information in the text?* Of the 47 meta-analyses acknowledging intercoder reliability by either generally discussing
intercoder reliability or by reporting intercoder reliability analyses (39.8% of the total sample), most meta-analyses (74.5%, or 35 meta-analyses) presented less than a paragraph about intercoder reliability in the text. Of the balance, 7 meta-analyses (14.9%) presented information about intercoder reliability in a separate section with “reliability” heading in the text, and 5 meta-analyses (10.6%) presented one or more complete paragraphs about reliability in the text.

What is the mean number of sentences devoted to the general discussion of intercoder reliability? Of meta-analyses that include a general discussion of intercoder reliability ($n = 38$, or 32.2% of the total sample), the mean number of sentences devoted to the general discussion of intercoder reliability was 1.5 ($SD = 0.95$). The number of sentences devoted to the general discussion of intercoder reliability ranged from 1 to 4 sentences. Of the 38 meta-analyses including a general discussion of intercoder reliability, 28 meta-analyses (73.7%) devoted one sentence, 4 meta-analyses (10.5%) devoted two sentences, 3 meta-analyses (7.9%) devoted three sentences, and 3 meta-analyses (7.9%) devoted four sentences to the reporting of intercoder reliability analyses. Further, most (92.1%, or 35 meta-analyses) of the 38 meta-analyses including a general discussion of intercoder reliability also reported intercoder reliability analyses.

What is the mean number of sentences devoted to the reporting of intercoder reliability analyses? Of meta-analyses reporting intercoder reliability statistics ($n = 44$, or 37.3% of the total sample), the mean number of sentences devoted to the reporting of intercoder reliability analyses was 1.5 ($SD = 0.82$). The number of sentences devoted to the reporting of intercoder reliability analyses ranged from 1 to 4 sentences. Of meta-analyses reporting intercoder reliability analyses, 29 meta-analyses (65.9%) devoted one
sentence, 12 meta-analyses (27.3%) devoted two sentences, and 3 meta-analyses (6.8%) devoted four sentences to the reporting of intercoder reliability analyses. Interestingly, 9 of the meta-analyses (20.5%) reporting intercoder reliability analyses did not include a general discussion of intercoder reliability.

**What percentage of meta-analyses include at least one citation about intercoder reliability?** Of the 118 meta-analyses in the study, 6 meta-analyses (5.1%) included at least one citation about intercoder reliability.

**What is the mean number of citations related to intercoder reliability?** Of meta-analyses including at least one citation about intercoder reliability (n = 6, or 5.1% of the total sample), the mean number of citations related to intercoder reliability analyses was 1.3 (SD = 0.5). The number of sentences devoted to the reporting of intercoder reliability analyses ranged from 1 to 2 citations. Four authors cited one reference and two authors cited two references related to intercoder reliability.


*What percentage of meta-analyses report a specific method for calculating intercoder reliability?* Of the 118 meta-analyses in the study, 44 meta-analyses (37.3%) reported a specific method for calculating intercoder reliability.

*What specific methods are being reported to calculate intercoder reliability?* Of the meta-analyses reporting a specific method for calculating intercoder reliability (*n* = 44, or 37.3% of the total sample), “intercoder agreement” was reported in 45.5% of the meta-analyses. Authors reported using “percentage agreement” in 40.9% of the meta-analyses. Responses provided are the exact description provided by the authors. Although “intercoder agreement” could indicate “Cohen’s kappa,” it could also “percentage agreement,” therefore the principal investigator did not collapse any categories where possible overlap could occur. Note that the percentages do not sum to 100% because some authors used more than one method to calculate intercoder reliability per meta-analysis. Table 21 presents a frequency distribution for the description of coders.
Table 21

*Specific Method Used to Calculate Intercoder Reliability*

<table>
<thead>
<tr>
<th>Reliability Method</th>
<th>Responses by Category</th>
<th>Percent</th>
<th>Total Cases&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohen’s kappa (or “kappa”)</td>
<td>6</td>
<td>11.5</td>
<td>13.6</td>
</tr>
<tr>
<td>Intercoder Agreement</td>
<td>20</td>
<td>38.5</td>
<td>45.5</td>
</tr>
<tr>
<td>Intercoder Reliability</td>
<td>1</td>
<td>1.9</td>
<td>2.3</td>
</tr>
<tr>
<td>Intraclass Correlation</td>
<td>1</td>
<td>1.9</td>
<td>2.3</td>
</tr>
<tr>
<td>Pearson’s r, Correlation Coefficient,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercoder Correlation</td>
<td>3</td>
<td>5.8</td>
<td>6.8</td>
</tr>
<tr>
<td>Percentage Agreement or Agreement Rate</td>
<td>18</td>
<td>34.6</td>
<td>40.9</td>
</tr>
<tr>
<td>Simple Agreement</td>
<td>1</td>
<td>1.9</td>
<td>2.3</td>
</tr>
<tr>
<td>Other</td>
<td>2</td>
<td>3.8</td>
<td>4.5</td>
</tr>
</tbody>
</table>

<sup>a</sup>Column does not sum to 100% due to multiple responses (n = 44).

What format are authors using to report reliability statistics? Of the authors reporting intercoder reliability analyses (n = 44, or 37.3% of the total sample), 42 meta-analyses (95.5%) reported reliability statistics in a text description only. The remaining two meta-analyses (4.5%) reported reliability statistics with a combination of text description and table(s).

If reliability statistics are not reported, and multiple coders were used, how do authors imply coder reliability? Coders did not respond to this item on the coding sheet. Generally speaking, study authors either reported intercoder reliability, or they did not.
Do authors differentiate between effect size coding and study descriptor coding?

Of the 44 meta-analyses in the study that reported intercoder reliability analyses, 10 meta-analyses (22.7%) differentiated between effect size coding and study descriptor coding.

Do authors report an overall/average intercoder reliability value for a group or groups of variables related to effect size in the study? Of the meta-analyses differentiating between effect size coding and study descriptor coding that reported intercoder reliability analyses \((n = 10)\), 8 meta-analyses (80%) reported an overall/average intercoder reliability value for a group or groups of variables related to effect size, and 2 meta-analyses (20%) did not report an overall/average reliability value.

Do authors report a range of intercoder reliability values for a group or groups of variables related to effect size in the study? Of the meta-analyses differentiating between effect size coding and study descriptor coding that reported intercoder reliability analyses \((n = 10)\), 1 meta-analysis (10%) reported a range of intercoder reliability values for a group or groups of variables related to effect size and 9 meta-analyses (90%) did not report a range of reliability values.

Do authors report specific intercoder reliability values for one or more variables related to effect size in the study? Of the meta-analyses differentiating between effect size coding and study descriptor coding that reported intercoder reliability analyses \((n = 10)\), 3 meta-analyses (30%) reported specific intercoder reliability values for one or more variables related to effect size, 7 meta-analyses (70%) did not report specific intercoder reliability values.
Do authors report intercoder reliability values for all variables related to effect size included in the study? Of the meta-analyses differentiating between effect size coding and study descriptor coding that reported intercoder reliability analyses (n = 10), 2 meta-analyses (20%) reported intercoder reliability values for all variables related to effect size and 8 meta-analyses (80%) did not report specific intercoder reliability values.

How clear (easy to find) was the information presented regarding the reporting of intercoder reliability of effect size variables? Of the meta-analyses differentiating between effect size coding and study descriptor coding that reported intercoder reliability analyses (n = 10), in all 10 of the meta-analyses, coders reported that “Yes, the information was clearly stated (easy to find).”

Do authors report an overall/average intercoder reliability value for a group or groups of variables related to study descriptors in the study? Of the meta-analyses differentiating between effect size coding and study descriptor coding that reported intercoder reliability analyses (n = 10), 7 meta-analyses (70%) reported an overall/average intercoder reliability value for a group or groups of variables related to study descriptors and 3 meta-analyses (30%) did not report an overall/average reliability value.

Do authors report a range of intercoder reliability values for a group or groups of variables related to study descriptors in the study? Of the meta-analyses differentiating between effect size coding and study descriptor coding that reported intercoder reliability analyses (n = 10), 3 meta-analyses (30%) reported a range of intercoder reliability values for a group or groups of variables related to study descriptors and 7 meta-analyses (70%) did not report a range of reliability values.
Do authors report specific intercoder reliability values for one or more variables related to study descriptors in the study? Of the meta-analyses differentiating between effect size coding and study descriptor coding that reported intercoder reliability analyses \((n = 10)\), 4 meta-analyses (40%) reported specific intercoder reliability values for one or more variables related to study descriptors and 6 meta-analyses (60%) did not report specific intercoder reliability values.

Do authors report intercoder reliability values for all variables related to study descriptors included in the study? Of the meta-analyses differentiating between effect size coding and study descriptor coding that reported intercoder reliability analyses \((n = 10)\), none of the meta-analyses reported intercoder reliability values for all variables related to study descriptors.

How clear (easy to find) was the information presented regarding the reporting of intercoder reliability of study descriptor variables? Of the meta-analyses differentiating between effect size coding and study descriptor coding that reported intercoder reliability analyses \((n = 10)\), in all 10 of the meta-analyses, coders reported that “Yes, the information was clearly stated (easy to find).”

If effect size and study descriptor variables are not differentiated, do authors report an overall/average intercoder reliability value for a group or groups of variables in the study? Of the meta-analyses that did not differentiate between effect size coding and study descriptor coding that reported intercoder reliability analyses \((n = 34)\), 31 meta-analyses (91.2%) reported an overall/average intercoder reliability value for a group or groups of variables and 3 meta-analyses (8.8%) did not report an overall/average reliability value.
If effect size and study descriptor variables are not differentiated, do authors report a range of intercoder reliability for a group or groups of variables in the study? Of the meta-analyses that did not differentiate between effect size coding and study descriptor coding that reported intercoder reliability analyses \((n = 34)\), 9 meta-analyses (26.5%) reported a range of intercoder reliability values for a group or groups of variables in the study and 25 meta-analyses (73.5%) did not report a range of intercoder reliability values.

If effect size and study descriptor variables are not differentiated, do authors report specific intercoder reliability values for one or more variables in the study? Of the meta-analyses that did not differentiate between effect size coding and study descriptor coding that reported intercoder reliability analyses \((n = 34)\), 21 meta-analyses (61.8%) reported specific intercoder reliability values for one or more variables in the study and 13 (38.2%) meta-analyses did not report an overall/average reliability value.

If effect size and study descriptor variables are not differentiated, do authors report intercoder reliability values for all variables included in the study? Of the meta-analyses that did not differentiate between effect size coding and study descriptor coding that reported intercoder reliability analyses \((n = 34)\), 2 meta-analyses (5.9%) reported specific intercoder reliability values for all variables included in the study and 32 meta-analyses (94.1%) did not report intercoder reliability values for all variables included in the study.

If effect size and study descriptor variables are not differentiated, how clear (easy to find) was the information presented regarding the reporting of intercoder reliability? Of the meta-analyses that did not differentiate between effect size coding and study descriptor coding that reported intercoder reliability analyses \((n = 34)\), 9 meta-analyses (26.5%) reported a range of intercoder reliability values for a group or groups of variables in the study and 25 meta-analyses (73.5%) did not report a range of intercoder reliability values.

If effect size and study descriptor variables are not differentiated, do authors report specific intercoder reliability values for one or more variables in the study? Of the meta-analyses that did not differentiate between effect size coding and study descriptor coding that reported intercoder reliability analyses \((n = 34)\), 21 meta-analyses (61.8%) reported specific intercoder reliability values for one or more variables in the study and 13 (38.2%) meta-analyses did not report an overall/average reliability value.

If effect size and study descriptor variables are not differentiated, do authors report intercoder reliability values for all variables included in the study? Of the meta-analyses that did not differentiate between effect size coding and study descriptor coding that reported intercoder reliability analyses \((n = 34)\), 2 meta-analyses (5.9%) reported specific intercoder reliability values for all variables included in the study and 32 meta-analyses (94.1%) did not report intercoder reliability values for all variables included in the study.

If effect size and study descriptor variables are not differentiated, how clear (easy to find) was the information presented regarding the reporting of intercoder reliability? Of the meta-analyses that did not differentiate between effect size coding and study descriptor coding that reported intercoder reliability analyses \((n = 34)\), 9 meta-analyses (26.5%) reported a range of intercoder reliability values for a group or groups of variables in the study and 25 meta-analyses (73.5%) did not report a range of intercoder reliability values.
descriptor coding that reported intercoder reliability analyses \( n = 34 \), in all 34 of the meta-analyses, coders reported that “Yes, the information was clearly stated (easy to find).”

*What percentage of meta-analyses include a discussion of “acceptable” reliability values?* Of the 118 meta-analyses in the study, 7 meta-analyses (5.9%) included a discussion of “acceptable” reliability values.

*What is the range of reported reliability values?* Table 22 presents the highest and lowest reliability criterions reported for each reported index of intercoder reliability. Responses provided are the exact description of the index reliability provided by the authors. Although “intercoder agreement” could indicate “Cohen’s kappa,” it could also “percentage agreement,” therefore the principal investigator did not collapse any categories where possible overlap could occur. Note that the table does not necessarily indicate matching pairs of data.
Table 22

*Range of Reported Intercoder Reliability Values*

<table>
<thead>
<tr>
<th>Reliability Method</th>
<th>High</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple Agreement</td>
<td>1.00</td>
<td>0.88</td>
</tr>
<tr>
<td>Cohen’s kappa or “kappa”</td>
<td>1.00</td>
<td>0.57</td>
</tr>
<tr>
<td>Intercoder/Interrater/Coder Agreement</td>
<td>1.00</td>
<td>0.33</td>
</tr>
<tr>
<td>Intercoder/Interrater Reliability</td>
<td>1.00</td>
<td>0.75</td>
</tr>
<tr>
<td>Intracoder Agreement</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Pearson’s r or Correlation</td>
<td>1.00</td>
<td>0.70</td>
</tr>
<tr>
<td>Percent Agreement/Agreement Rate</td>
<td>1.00</td>
<td>0.79</td>
</tr>
<tr>
<td>Other</td>
<td>1.00</td>
<td>0.87</td>
</tr>
</tbody>
</table>

*Note.* The values reported do not necessarily represent matching pairs of data.

How clear (easy to find) was the information presented regarding the reporting of reliability values? Of the 44 meta-analyses in the study that reported intercoder reliability analyses, coders reported that “Yes, the information was clearly stated (easy to find)” for 35 meta-analyses (79.5%). For the remaining nine meta-analyses (20.5%), coders reported that “No, the information was not clear (easy to find). I had to make assumptions to code this variable.”

What is the lowest reliability criterion deemed “acceptable” by authors? As authors of studies did not report excluding any variables based on reliability values, Table
22 also presents the lowest accepted values deemed “acceptable” by authors for each reported index of reliability.
CHAPTER FIVE

Summary, Conclusions, and Recommendations

This chapter consists of six main sections. The first section contains a summary of the present study including the statement of the problem and methodology used for research. The second section contains a summary of the findings. This section includes a summary of the interrater reliability analyses and a summary of the 48 sub-problems related to the overarching research question. The third section addresses the conclusions based on the findings. This section includes conclusions based upon the study methodology and the research findings. The fourth section presents a series of recommendations. The sixth section addresses limitations of the study. The final section addresses future practical matters.

Summary of the Study

In both quantitative and qualitative research methods, there are criteria for judging the objectivity and credibility of the data collection and measurement processes (Light, Singer, & Willett, 1990; Patton, 2002). The meta-analyst is not exempt from addressing these measurement issues and must carefully consider the standards of “objectivity, systematicity, and rigor” in order to insure the credibility of the research (Cooper & Hedges, 1994, p. 13). Researchers tend to agree that meta-analysts should “provide a rigorous assessment of interrater reliability,” but the literature pertaining to how to select a reliability coefficient, what determines an acceptable level of reliability, and the criteria for the reporting of coding procedures and intercoder reliability is often very general and left to the researcher for interpretation (Bullock & Svyantek, 1985, p. 114). Therefore, the
idea of credibility in meta-analysis, specifically as it relates to intercoder reliability, is the focus of this work.

Statement of the Problem

This study examined 48 sub-problems in order to provide a detailed response to the research question: How is intercoder reliability assessed and reported in published meta-analyses related to preschool through grade 12 education? The sub-problems did not represent a set of criteria for judging the quality of intercoder reliability reporting in meta-analyses. They simply served as a tool to explore the current state of intercoder reliability reporting among published meta-analyses related to preschool through grade 12 education. The subproblems were divided into seven categories: the use of meta-analysis in education, coders and coding strategy, coder training procedures, the code book, addressing intercoder reliability, and reporting intercoder reliability.

Methodology

The initial data set for this study included 100 journal articles acquired from the three databases: ERIC, Education Abstracts, and PsychInfo. Journal articles were included in the final sample based on the following inclusion criteria: (a) journal articles that employ meta-analytic methods and report meta-analytic results, and (b) meta-analyses covering content related to education in preschool through grade 12. Further, journal articles that were methodological and theoretical in nature with no reported meta-analytic results were not eligible for inclusion the study. During the coding process, one article did not meet inclusion criteria and was excluded from analysis. Within the remaining 99 journal articles comprising the final data set, 10 articles reported results from more than one meta-analysis. Each meta-analysis (whether the author reported
In the current study, there were two coders, each coding all studies. The coders for this study were trained undergraduate students. Both coders were candidates for graduation with knowledge of research methods. Coder training was based on Stock’s (1994, p. 134-135) eight steps for training coders participating in a meta-analysis. Three coding meetings took place over a one-month period. Each meeting lasted approximately two hours. Coders resolved coding discrepancies that occurred in the reliability samples by consensus. To reach consensus, the principal investigator assisted by directing the coders to the code book for guidance and facilitating a metacognitive discussion about how to code a particular item. After such discussion, coders always reached consensus.

The coding instrument for this study utilized items based on variables coded in two previous studies: Lombard, Snyder-Duch, and Bracken’s (2002) study that assessed the reporting of intercoder reliability in content analyses in the mass communications literature and Steiner et al.’s (1991) study which reviewed meta-analyses in organizational behavior and human resources management.

To establish interrater reliability for the coding instrument, the study used a four-stage reliability analysis. The design of the reliability analysis, based on the original data set, included 100 journal articles. The initial reliability analysis, which consisted of 20% of the journal articles determined eligible for inclusion, established baseline interrater reliability. After the initial reliability analysis, the principal investigator split the remaining journal articles into two samples (containing 40 articles each) to examine interrater reliability over time. The final sample, first half reliability analysis consisted of
eight articles (20% of the sample). The final sample, second half reliability analysis also consisted of eight articles (20% of the sample). The overall reliability analysis, used to establish overall interrater reliability, consisted of all 36 articles utilized in the previous three reliability analyses. Thus, the overall reliability analysis accounted for 36.4% of the total articles ($N = 99$) included in the study.

This study used an item-by-item approach for all four reliability analyses (Orwin & Corday, 1985; Orwin, 1994). The study employed a design 2 generalizability theory approach to assess interrater reliability because the study utilized two coders, each coder rating all studies in the reliability sample (Crocker & Algina, 1986). Therefore, the appropriate ANOVA model for this study was a two-way mixed model. It follows that, for categorical variables, the appropriate generalizability coefficient was a design 2 generalizability coefficient with an absolute decision (ICC[A,k]). For quantitative variables, the appropriate generalizability coefficient was a design 2 generalizability coefficient with a relative decision (ICC[C,k]) (McGraw & Wong, 1996; Nichols, 1999).

However, after examining the distribution of the data, it became apparent that choosing an alternate reliability index for some instrument items was essential. Due to its ease of calculation and interpretation, the study used percentage agreement as a secondary index of interrater reliability.

In the meta-analysis literature, there are no established standards for determining what signifies an acceptable level of reliability. In this study, the principal investigator used Neuendorf’s (2002) conventional rule that reliability coefficients of .80 or greater are acceptable in “most situations” (p.143) as a baseline. However, it is difficult to use a conventional rule that applies to reliability coefficients when each reliability coefficient
has a separate mathematical basis and purpose. Therefore, the principal investigator used reliability coefficients as a method of screening data to assess for potential issues with data integrity.

The principal investigator used SPSS software for all data analyses. In addition to the reliability analysis described previously, the principal investigator used descriptive statistics to address the overarching research question and the related 48 sub-problems.

*Summary of the Findings*

*Interrater Reliability*

In this study, the principal investigator used Neuendorf’s (2002) conventional rule that reliability coefficients of .80 or greater are acceptable in “most situations” (p.143) as a baseline. For the initial reliability analyses, ICC(A,k) and ICC(C,k) were almost all within the guideline. For the final sample, first half reliability analyses, ICC(A,k) and ICC(C,k) were well within the guideline. For the final sample, second half analyses, ICC(A,k) and ICC(C,k) fell outside the guideline, suggesting that coder fatigue was a factor in this study. As for the overall reliability sample, ICC(A,k) and ICC(C,k) were well within the guideline.

For the items where percentage agreement was the most appropriate index of reliability, there were cases in all four reliability samples where interrater reliability equaled zero. Because interrater reliability was assessed on an item-by-item basis, the principal investigator was able explore items having poor interrater reliability. In general, items for which percentage agreement equaled zero were cases when the sample size was very small. For example, there was one case in the sample and the coders disagreed on the coding of that particular case.
The principal investigator used the interrater reliability coefficients to screen the final data set. The use of the item-by-item approach to interrater reliability allowed the principal investigator to do this by examining unreliable items one a case-by-case basis and make decisions on how to handle questionable items in data analysis.

Sub-problems

The subproblems were divided into six categories: the use of meta-analysis in education, coders and coding strategy, coder training procedures, the code book, addressing intercoder reliability, and reporting intercoder reliability.

Language arts education was the content area that had the most published meta-analyses. The next most represented content areas were special education and academic (educational) achievement followed by arts education.

Authors reported the number of coders less than half of the time. Authors reported a description of the coders’ levels of expertise less than a quarter of the time. Of the meta-analyses reporting a description of the coders’ levels of expertise, over half of the meta-analyses described coders as “study authors.” Authors rarely reported coder training procedures. Authors also rarely reported using a criterion for consistency (e.g., stopping rule) to determine that coders were adequately trained.

Just under half of the authors provided an extended description of the code book used in the study. Less than 10% of authors state that a copy of the code book (or coding protocol) is available upon request.

Authors differentiated between the reliability sample and the actual sample less than half of the time. Authors acknowledged intercoder reliability less than half of the time. Authors rarely cite or reference other authors regarding intercoder reliability.
Authors reported intercoder reliability analyses less than half of the time. Authors reported a specific method for calculating intercoder reliability just over a third of the time. Of the meta-analyses reporting a specific method for calculating intercoder reliability, “intercoder agreement” was reported in 45.5% of the meta-analyses. Authors reported using “percentage agreement” in 40.9% of the meta-analyses.

In general, authors are not differentiating between effect size coding and study descriptor coding when assessing intercoder reliability. Of the meta-analyses that did not differentiate between effect size coding and study descriptor coding and reported intercoder reliability analyses, most meta-analyses reported an overall or average intercoder reliability value for a group or groups of variables.

Authors rarely included a discussion of “acceptable” reliability values. The range of reliability values reported for each index of reliability is as follows: agreement rate/simple agreement (0.88 – 1.00), Cohen’s kappa or “kappa” (0.57 – 1.00), intercoder/interrater/coder agreement (0.33 – 1.00), intercoder/interrater reliability (0.75 – 1.00), intracoder agreement (1.00, only value reported), Pearson’s r or correlation (0.70 – 1.00), and percent agreement (0.79 – 1.00).

Conclusions

This study was a content analysis of meta-analyses published related to preschool through grade 12 education. Because the coding process in content analysis is similar to the coding process in meta-analysis, the principal investigator was able to draw conclusions from both the study methodology and the research findings. Therefore, this section outlines conclusions based upon the study methodology and the research findings.
Study Methodology

This section presents conclusions drawn from the study methodology based upon the following categories: coders and coding strategy, coder training procedures, the code book, and assessing and reporting intercoder reliability.

Coders and coding strategy. It is important for the coders to have input in designing the coding strategy. For example, in this study, the coders requested that the principal investigator split the final sample into two parts in order for them to maintain motivation and stay attentive during the coding process. Coders are investing a substantial amount of time and effort into a project of this magnitude and project leadership should carefully consider any adjustments or suggestions made by the coders that do not affect the outcome of the study. Investigators should also be aware that coder fatigue is a real threat to coder reliability. By using the four-stage reliability assessment protocol, the principal investigator was able to determine that coder fatigue was potentially a factor in this study.

Early in the research design phase, principal investigators should carefully outline a plan for assessing interrater reliability. Although interrater reliability assessment is not the main analysis in a meta-analysis, it helps insure that data collection procedures are both credible and efficient. Of meta-analyses reporting the number of primary studies included in the reliability sample and the total number of studies in the meta-analysis ($n = 47$), 68.1% of the meta-analysis utilized the entire sample as the reliability sample. This defeats the purpose of utilizing a reliability sample, as double coding all studies in a meta-analysis is a time-consuming and daunting task. With respect to the size of the
reliability sample, both the coders and the principal investigator for this project exceeded the 10 - 20% guideline presented by Wimmer and Dominick (2003).

*Coder training procedures.* This project had three coding meetings. The first coding meeting served as a coder training meeting where the principal investigator outlined the purpose of the study and the task at hand. During this meeting, it is essential for the principal investigator to use a team-building approach. A team-building approach invests the coders in the study and establishes them as equal partners in the coding portion of the project.

Resolving coder discrepancies is an essential part of the reliability analysis, as the principal investigator must reduce the two data sets for each article in the reliability sample to a single data set. In this study, the primary purpose of the second coding meeting was to resolve coder discrepancies that occurred during the initial reliability sample. By using a metacognitive process and encouraging dialogue between coders, coders were able not only to resolve discrepancies, but also set precedents for coding decisions in the final sample. This created more consistency in coding decisions, as coder reliability, in general, increased from the initial reliability analysis to the final sample, first half reliability analysis.

*Code book.* Providing the coders with a sample study and copies of the code book and coding sheet prior to the initial training session was very valuable. At the initial coding meeting, or training session, coders were able to raise questions or concerns and discuss any ambiguities that may have existed with the coding sheet. Typically, the principal investigator designs coding sheets in a manner that follows his/her thought process. Because the coders are doing the actual coding of the study, it is important that
the coding sheet be compatible with their thought processes. Therefore, project leadership should carefully consider any adjustments or suggestions regarding the code book made by the coders that do not affect the outcome of the study.

Assessing and reporting intercoder reliability. This study utilized an item-by-item approach to assessing interrater reliability. The item-by-item approach was very effective in helping the principal investigator to identify flawed items or items where the coders were not attending to the task. For example, item one asked coders to identify the content area of the meta-analysis. In general, interrater reliability was high (ICC[A,k] for the overall reliability sample = 0.99) for the coders’ first choice of content area. However, coders were able to select more than one content area per item. As coders expressed their second, third, and fourth choices, interrater reliability dropped dramatically, with percent agreement equaling zero in some cases. Upon examination of the raw data, it was clear that one coder was consistently listing the “other” category (being overly thorough), while the other coder was simply stopping at the main content areas. Therefore, in the final data set, the principal investigator eliminated “other” responses – except when there was no other content area listed.

An example of coders not attending to the task was with items 7e to 7u. These items related to the reporting of intercoder reliability. Generally, coders responded to these items in one of three ways, “yes,” “no,” or “standardized reliability criterion values not reported.” If the meta-analysis did not report standardized reliability criterion values, then coders should respond as such throughout the entire strand of items. However, there were several occasions where coders were not attending to the task and responded to some items as “no” instead of “standardized reliability criterion values not reported.”
This impossible coding pattern occurred throughout the coding process and peaked in some instances in the final sample, second half reliability sample, indicating that coder fatigue had become a factor. By assessing interrater reliability on an item-by-item basis, it was easy for the principal investigator to identify the problem and thereby, identify the impossible coding pattern in the final data set and screen the final data set as necessary.

The interrater reliability analyses used a design 2 generalizability theory approach, which resulted in the calculation of design 2 generalizability coefficients for both norm-referenced (ICC[C,k]) and criterion-referenced (ICC[A,k]). Combining the item-by-item and generalizability theory approaches assisted the principal investigator in carefully planning the design of the interrater reliability analyses. By combining these approaches, the principal investigator was able to calculate separate reliability coefficients, based on appropriate estimators, individual coding items.

It is important to note that the design 2 generalizability coefficients were not appropriate for all variables in the study. As with all statistical methodologies, there are caveats associated with generalizability theory coefficients. Because generalizability coefficients (and their corresponding intraclass correlations) utilize an ANOVA-based approach, there must be substantial between coder variance to produce a substantial indication of agreement (Selvage, 1976). As such, generalizability coefficients (and their corresponding intraclass correlations) are not appropriate when the distribution of the ratings is concentrated in a small range (Jones et al., 1983).

Although a substantial number of items in this study used percentage agreement in lieu of generalizability coefficients due to a lack of between coder variance, it is likely that phenomenon occurred as an artifact of the data, due to the nature of the coding sheet,
rather than a lack of applicability of the method. Most of the meta-analyses did not report the reliability information, resulting in multiple cases where a large percentage of coder responses were “no” or “standardized reliability not reported.” In cases where coders are agreeing, using the same response for all items, generalizability coefficients (and their corresponding intraclass correlations) are not appropriate because of the corresponding lack of between coder variance. By using the item-by-item and generalizability theory approaches in conjunction, the principal investigator was able to screen for items where percentage agreement was a more appropriate index of interrater reliability than generalizability theory.

It was easy for the principal investigator to calculate both indices of interrater reliability. SPSS, a statistical package familiar to most researchers, calculates generalizability coefficients via the intraclass correlation option. Percentage agreement, which is easily calculated by hand in small data sets, can also be calculated in SPSS using the macro available free of charge from Lombard (2005).

Research Findings

This section presents conclusions drawn from the research findings. In the study sample, language arts education, special education, arts education, and academic (educational) achievement had the most published meta-analyses. This finding is consistent with the fact that within the educational literature, special education researchers appear to be leading the charge with respect to methodological research related to meta-analysis and meta-analysis reporting. Mostert’s (1996) work examines the reporting for meta-analyses in learning disabilities research. Kavale (2001), a faculty member from the Department of Special Education at the University of Iowa, has
published a primer on the methods of meta-analysis for educational researchers. Further, Forness (2001), who conducted a mega-analysis of 24 meta-analyses related to special education, has maintained that the development of best practice in special education relies on the “accumulation of research findings” and that meta-analysis is an important tool to “synthesize cumulative research findings on specific interventions in, or related to, special education” (p. 185).

Most authors did not report the number of coders or the levels of the coders’ expertise. In those cases where the authors reported the number of coders, most authors reported using two coders. In those cases where authors provided a description of the coders, authors most frequently identified the coders as “study authors.” Bias is a potential threat to data credibility when study authors code the primary studies in the meta-analysis.

Most authors do not differentiate between the reliability sample and the actual sample. Of those reporting a reliability sample, most authors use the entire sample as the reliability sample. This implies that authors, in general, are double coding all studies in their meta-analyses. Double coding is occurring in both small-scale and large-scale studies. With techniques available to assess intercoder reliability, there is no need for double coding in large-scale studies. Using a well-designed methodology to assess intercoder reliability can save time and resources that are expended when entire samples of primary studies are double coded.

Most authors do not report any details related to coder training or description of their methods to resolve coding discrepancies. Further, most authors acknowledge the use of a code book, but few offer much detail regarding the variables in the code book. This
lack of detail in reporting could be due to two reasons: (a) authors do not think that this information is important and valuable to report or (b) journal editors are removing such detail from reports due to space constraints.

Authors report assessing intercoder reliability just over a third of the time. This is consistent with the findings of Steiner et al. (1991), who reviewed meta-analyses in organizational behavior and human resources management and found that only 9 of 35 studies (25.7%) reported interrater agreement for any coding they did. It is also consistent with the research of Lombard, Snyder-Duch, and Bracken (2002), who analyzed 200 content analyses in the mass communications literature and find that only 69% of the content analyses contained any report of intercoder reliability. With meta-analysis becoming a widely accepted methodology, it is important for authors to address the credibility of their data. In both quantitative and qualitative research methods, there are criteria for judging the objectivity and credibility of the data collection and measurement processes (Light, Singer, & Willett, 1990; Patton, 2002). The meta-analyst is not exempt from addressing these measurement issues. Assessment of intercoder reliability is important because a coding sheet is a research instrument and therefore is subject to the same rigorous standards as instrumentation in primary research.

**Recommendations**

This section offers recommendations for best practice in interrater reliability analysis and reporting based on the study methodology and the research findings. This section also makes recommendations for future research.
Methodology

The principal investigator recommends that meta-analysts utilize an item-by-item approach to assessing interrater reliability. This approach gives meta-analysts the ability to isolate “problem” variables and either revise the coding protocol to improve reliability or screen the final data set for impossible data patterns. Further, to prevent impossible data patterns from occurring, the principal investigator recommends noting such patterns during coder training meetings and in the coding manual and encouraging coders to look for such patterns as a way to double-check their work.

The principal investigator recommends that meta-analysts carefully consider their choice of reliability coefficient. When choosing a reliability coefficient, it helps to have an overarching design for the reliability analysis. When meta-analysts design a reliability analysis at the start of a project, it helps them not only to make sound, justifiable decisions but also to anticipate potential problems. This process is akin to primary researchers thinking through a research design prior to the start of a study.

The principal investigator recommends that meta-analysts use the generalizability theory framework, in conjunction with the item-by-item approach, to design interrater reliability analyses. Using this design, the meta-analyst is required to carefully think through the reliability analysis and then select the appropriate generalizability coefficient (and corresponding intraclass correlation) for each variable. This design is very flexible as it accommodates both categorical and quantitative variables and both criterion- and norm-referenced decisions. This design is easy to calculate using a widely distributed statistical package, SPSS. Finally, this design is easy to interpret.
The principal investigator recommends that meta-analysts understand that generalizability theory, like many theories, has its limitations. Therefore, meta-analysts choosing a generalizability framework should also have an alternate framework for assessing interrater reliability.

The principal investigator recommends that meta-analyst use, at minimum, percentage agreement as an alternate framework for assessing interrater reliability. Although percentage agreement does not account for chance agreement (like Cohen’s kappa), it is a good initial screen for data integrity. Meta-analysts can often use percentage in situations where generalizability theory is not appropriate. Percentage agreement is easy to calculate – by hand or using SPSS. Further, with so many meta-analysts failing to assess intercoder reliability, percentage agreement is easy and intuitive to interpret, which may make meta-analysts more likely to use this method rather than other indices of interrater reliability. Perhaps a better alternate framework would be using Cohen’s kappa in conjunction with percentage agreement. By examining overall agreement in terms of a percentage and examining agreement while accounting for agreement due to chance, meta-analysts will have a more complete picture of data integrity.

The principal investigator recommends that meta-analysts carefully consider the hierarchical nature of their coding items and decide whether calculating agreement is appropriate for all levels of an item. For example, in this study, item one asked coders to identify the content area of the meta-analysis. Coders could choose as many content areas as they wished, numbering their first choice with “1” and so on. In general, coder reliability was high for the coders’ first choice of content area. However, as coders made
second, third, and in some cases, fourth decisions, intercoder reliability dropped dramatically. Meta-analysts should ask the question “Is it reasonable to expect that coders agree the further down the hierarchy of choices one explores?” In most cases, it is reasonable to expect strong agreement for the first choice, but the subsequent choices are often a matter of fine distinction and strong agreement is not necessarily expected or necessary.

Research Findings

The principal investigator recommends that meta-analysts address the credibility of their data. This helps to avoid the “garbage in, garbage out” phenomenon that can occur in research of any kind. If more and more meta-analysts address the credibility of their data, it enhances the credibility of meta-analysis as a research method.

The principal investigator recommends that meta-analysts acknowledge that the establishment of interrater reliability is necessary any time a coding instrument is used. Just as in primary research, reliability (of any kind) is not inherited; it must be established every time an instrument is used.

The principal investigator recommends that meta-analysts report the following details regarding the coding process: number of coders, description of coders’ levels of expertise, coder training procedures, and the method for resolving coding discrepancies. By reporting this information, the meta-analyst establishes to the readership that data integrity and credibility was a priority for the study.

The principal investigator recommends that meta-analysts report the following details regarding the coding instrument: how the instrument was developed, and how interrater reliability was established. Again, by reporting this information, the meta-
analyst establishes to the readership that data integrity and credibility was a priority for the study. Further, to encourage replication of research, the principal investigator recommends that meta-analysts make instrumentation available to other researchers upon request.

The principal investigator recommends that meta-analysts report the following details regarding the interrater reliability analysis: reliability sample information and method(s) used to establish interrater reliability. The principal investigator recommends that meta-analysts report reliability coefficients in one or more of the following formats: (a) specific reliability coefficients for all variables involved in the study, this could be easily added to tables reporting descriptive statistics; (b) range of reliability coefficients for each index of reliability used; or (c) means and standard deviations of reliability coefficients for each index of reliability used. By being transparent about the establishment of interrater reliability, meta-analysts enhance the credibility and integrity of their data.

Future Research

The principal investigator recommends that methodological research related to meta-analysis focus on Monte Carlo simulations to determine the impact of coder unreliability on the outcomes of a meta-analysis. To date, there have been no studies that have assessed the impact of unreliable data on the outcomes of a meta-analysis. If journal editors are going to require the reporting of interrater reliability analyses, it will most likely require empirical evidence to spark such a requirement.

The principal investigator recommends that methodological research related to meta-analysis focus on Monte Carlo simulations to determine guidelines for an
“acceptable” level of reliability for each different reliability index. Because interrater reliability coefficients are mathematically different and indices have different assumptions and interpretations, conventional rules must be coefficient specific.

The principal investigator recommends that methodological research related to intercoder reliability focus on determining the relevance for setting standards for interrater reliability in meta-analysis or content analysis by using confidence intervals to establish a desired degree of accuracy for individual scores and then using the formula for standard error to estimate the “needed” reliability.

The principal investigator recommends the expansion of this study to include a by-journal analysis. This would allow researchers to determine if certain journals are requiring a more stringent level of reporting.

The principal investigator recommends a qualitative research study that examines the coding process, from both the principal investigator’s and the coders’ points of view. Examination of the coding process would allow researchers to develop a better understanding of the process, highlighting both strengths and barriers.

Limitations

The primary limitation of this study was potential intercoder unreliability. Although the study employed a carefully designed coding process and reliability assessment protocol, there were variables where intercoder reliability was an issue.

A secondary limitation of this study was the use of undergraduate students as coders. Although the content analysis revealed little information regarding the expertise of coders, none of the meta-analyses in the study reported using undergraduate students as coders.
Because this study was limited to published reports, it was not exempt from availability bias (also known as publication, retrieval, or selection bias), an issue often faced by meta-analysts. However, because this study was descriptive in nature, the principal investigator simply acknowledged the availability bias that is inherent to this type of research.

Finally, a limitation to this study was the inclusion of more than one meta-analysis per journal article. The overall data set included 99 journal articles. Within these journal articles, 10 articles reported results from more than one meta-analysis. Because some authors reported interrater reliability information for more than one meta-analysis, each meta-analysis (whether the author reported interrater reliability information or not) was considered a separate case. Therefore, the final data set was comprised of 118 meta-analyses reported in 99 journal articles. By including more than one meta-analysis per journal article, this could potentially influence the descriptive statistics because some articles were weighed by a factor of how many meta-analyses were reported in the article.

**Future Practical Matters**

In order for the reporting of intercoder reliability in meta-analyses to change, journal editors must recognize the value of such reporting and require authors to report such information. For this to occur, journal editors must examine the empirical evidence with respect to data credibility in meta-analysis and then allocate space within their journals for meta-analysts to address measurement issues. Further, journal editors should work with meta-analysts to establish a consistent reporting format for the reporting of intercoder reliability.
Currently, traditional software packages such as SPSS cannot calculate most interrater reliability indices. In order for it to become commonplace for the meta-analyst to assess and report intercoder reliability, traditional software packages must have the capability to calculate intercoder reliability with an easy-to-use interface.
REFERENCES


BIBLIOGRAPHY OF JOURNAL ARTICLES USED IN CONTENT ANALYSIS


Appendix A
Assessment and Reporting of Intercoder Reliability in Published Meta-Analyses
Coding Manual
Spring 2006

Coder: _____________________________  Date Coded: _____________________________
Article Number: ____________________  Date Entered: ____________________________

1. What is the content area of the meta-analysis? (Please rank your decisions with “1” indicating the most appropriate choice(s). You may choose multiple content areas if applicable. You may assign a decision rank to more than one category, i.e., you may indicate that two content areas are the most appropriate choice by assigning them BOTH a rank of “1”.)

_____ Academic (Educational) Achievement
_____ Arts Education
_____ Computer-Assisted Instruction
_____ Consumer Sciences
_____ English as a Second Language
_____ Elementary School Leadership
_____ Foreign Language Education
_____ General Early Childhood Education
_____ General Elementary Education
_____ General Middle & High School Education
_____ Health Education
_____ Language Arts Education
_____ Mathematics Education
_____ Middle & High School Leadership
_____ Physical Education
_____ Science Education
_____ Social Studies Education
_____ Special Education
_____ School Psychology/Guidance Counseling
_____ Technical Education
_____ Vocational Education
_____ Other: ____________________________

DEFINITIONS:
Academic (Educational) Achievement- Analysis of educational research related to academic achievement or educational attainment of students in preschool – grade 12.

Arts Education- Analysis of educational research related to the teaching and learning of the fine arts: e.g., art, music, drama, band/ orchestra, choir, etc.
Computer-Assisted Instruction- Analysis of educational research related to the use of computer technology as a mode of instruction

Consumer Sciences Education- Analysis of educational research related to the teaching and learning of home economics, family studies, human relations, etc.

English as a Second Language- Analysis of educational research related to the teaching and learning of English as a second language (ESL)

Elementary School Leadership- Analysis of educational research related to leadership of elementary (K – 6th grade) schools: e.g., principals, administration, etc.

Foreign Language Education- Analysis of educational research related to the teaching and learning of foreign languages: e.g., French, Spanish, Italian, Latin, German, etc.
General Early Childhood Education- Analysis of educational research related to preschool-age children, day care, infant education, toddler education, etc. that does not focus on a specific content area.

General Elementary Education- Analysis of educational research related to the education of elementary school-aged (K-6th grade) students that does not focus on a specific content area.

General Middle & High School Education- Analysis of educational research related to the education of middle and high school-aged (7th – 12th grade) students that does not focus on a specific content area.

Health Education- Analysis of educational research related to the teaching and learning of health education

Language Arts Education- Analysis of educational research related to the teaching and learning of language arts: e.g., English, literature, grammar, reading, etc.

Mathematics Education- Analysis of educational research related to the teaching and learning of mathematics: e.g., general math, algebra, geometry, trigonometry, calculus, business math, etc.
Middle & High School Leadership- Analysis of educational research related to leadership of middle and high (7th – 12th grade) schools: e.g., principals, administration, etc.

Physical Education- Analysis of educational research related to physical education: e.g., physical activity, gym, etc.

Science Education- Analysis of educational research related to the teaching and learning of science: e.g., general science, chemistry, biology, anatomy, botany, etc.

Social Studies Education- Analysis of educational research related to the teaching and learning of social studies: e.g., civics, history, geography, government, etc.

Special Education- Analysis of educational research related to special education, the teaching of special education, or learning of special needs students not related to a specific content area: e.g., learning disabilities, mental retardation, special needs, mild-to-moderate special needs, intensive special needs, specific conditions (Autism, Down’s syndrome, etc.)

School Psychology/Guidance Counseling- Analysis of educational research related to the provision and delivery of school psychology and guidance counseling services to children in the school setting

Technical Education- Analysis of educational research related to the teaching and learning of technical education: e.g., graphic design, metal shop, wood shop, computer technology, etc.
Vocational Education- Analysis of educational research related to the teaching and learning of vocational education: e.g., cosmetology, auto body, hospitality, etc.

Other- Any other term or terms used to describe the content area of the meta-analysis not listed above.

2. **How many primary studies are included in the meta-analysis?**

**DEFINITION:**
Primary studies- The primary research studies in the text that are included in the meta-analysis.

3. **CODERS and CODING STRATEGY**

3a. **Did the author report the number of coders involved in coding the primary studies?**

- Yes
- No

**DEFINITION:**
Coder- A researcher who locates, reduces, manipulates, and transcribes information.

3b. **How many coders were involved in coding the primary studies?**

(Check all that apply.)

- One coder
- More than one coder (List specific number of coders: ________________, if number of coders is not reported, enter NR.)
- The author does not report any details about the person(s) involved in the coding process.

**NOTES:**
If coders coded in pairs/coded jointly, code the total number of people who coded, NOT the number of pairs.
If reliability is reported, there must be more than one coder. Therefore, if a definite number is not given, select “more than one coder.”

If the text uses the word "Coders" (plural) or the phrase "we coded/analyzed/examined/etc.,” but does not identify a specific number of coders, then select "more than one coder," and enter “NR” for the number of coders reported.

If an individual steps in ONLY to resolve or consult on disagreements, he or she should not be considered a coder.

If the text does not use the word “coder” or “coders” or any other term to indicate the individuals coding the primary studies, then select “the author does not report any details about the person(s) involved in the coding process.”
3c. Was the information needed to code variable 3b clearly stated (easy to find) in the text?

- Yes, the information was clearly stated (easy to find).
- No, the information was not clear. I had to make assumptions to code this variable.
- I was unable to ascertain any information about the individuals involved in coding the primary studies.

Notes:
If the author clearly states the number of coders as a numeric value, code this variable as “yes.”

If it is necessary to refer to how many authors there are to code this variable because the text states that “Each co-author coded X number of units,” code as “yes, the information was clearly stated (easy to find).”

If it is necessary to do any calculations, look in tables or footnotes, or make any assumptions to code this variable, code this variable as “no.”

In all cases in which “more than one” is selected and “NR” (NOT REPORTED) entered, code this variable as "no."

If you selected “the author does not report any details about the person(s) involved in the coding process” for variable 3b, code this variable as “no.”

3d. Did the author(s) differentiate between coding of the reliability sample and coding of the actual sample?

- Yes
- No

Definition:
Reliability sample – primary studies that are coded in order to assess reliability; this may be a separate set of primary studies, a portion of the actual sample, or all primary studies in the actual sample.

Actual sample – primary studies that are coded and included in the final meta-analysis.

3e. What was the size of the reliability sample?

- Number of primary studies in the reliability sample (You may enter “0.” You may enter the overall sample size if appropriate.)
- The authors did not report any information regarding the assessment of reliability.
NOTES:
If a CLEAR percentage of the overall sample is given (e.g., “To assess reliability, 20% of the total sample was coded by all coders”), calculate and list the reliability sample size. If the calculation does not come to a whole number, round to the nearest whole number (i.e., round up for .5 or higher, round down for numbers below .5). If the text gives a vague percentage (e.g., “More than 20%...”), then code as not reported.

If the text states that the coders coded "the data" and does not indicate that there was a separate reliability sample, then assume that the reliability sample size is the same as the overall sample size.

Read carefully when a text mentions a "pilot study;" a text may refer to a "pilot study" and actually be describing a reliability test.

If more than one reliability test was performed, report the reliability sample size of the latest test of reliability in the study.

If the authors do not report any information regarding the assessment of reliability, select “the authors did not report any information regarding the assessment of reliability.”

3f. Was the information needed to code variable 3e clearly stated (easy to find) in the text?

______Yes, the information was clearly stated (easy to find).

______No, the information was not clear. I had to make assumptions to code this variable.

______The authors did not report any information regarding the assessment of reliability.

NOTE:
If it is necessary to do any calculations, look in tables or footnotes, or make any assumptions to code this variable, select the "No" option.

If you selected “the authors did not report any information regarding the assessment of reliability” for variable 3e, select it again for variable 3f.

3g. What proportion of the primary studies were included in the reliability sample?

______Total number of primary studies coded (See variable 2.)

______Total number of primary studies included in the reliability sample (See variable 3e.)

______Proportion of primary studies coded by multiple coders (Total number of studies coded by multiple coders divided by total number of studies coded.)

______The author does not present enough information in the text to code this variable.

NOTE:
If the calculation does not come to a whole number, round to the nearest whole number (i.e., round up for .5 or higher, round down for numbers below .5).
3h. Did the author provide a description of all of the coders’ levels of expertise?

_____ Yes

_____ No

DEFINITION:
Level of expertise- Coders educational background: e.g., undergraduate research assistants, graduate research assistants, study authors, etc.

3i. How did the author describe the coders’ levels of expertise?

_____ Graduate students/graduate research assistants

_____ Undergraduate students/undergraduate research assistants

_____ Study authors

_____ Other: ________________________________

DEFINITIONS:
Graduate students/graduate research assistants- individuals enrolled in a graduate program assisting in the research

Undergraduate students/undergraduate research assistants- individuals enrolled in an undergraduate program assisting in the research

Study authors- the actual authors of the meta-analysis

Other- Any other term or terms used to describe the coders not listed above.

4. CODER TRAINING

4a. Does the text state how coders were trained?

_____ Yes

_____ No

DEFINITION:
Coder training- The process used to make certain that coders are adhering to the code book and utilizing the code book and/or coding sheet(s) in a uniform manner.
4b. Amount of training (fill in amount; then check unit):

Number: ____________ of

______ Hours

______ Units

______ Sessions

______ OTHER _________________________________________

______ AMOUNT OF TRAINING NOT REPORTED

NOTES:
Be as specific as possible in coding the amount of training. For example, if the text says that
coders were trained during "two 2-hour sessions," code as 4 hours.

"Descriptors" include such references as "several," "a few," "many." List a descriptor only if a
specific number is not provided.

To merely state that coders were "trained" is not enough to be coded here, check "AMOUNT OF
TRAINING NOT REPORTED."

4c. How does the author describe the coder training procedure? (Check one.)

______ 1 – Coder training is kept secret; not described at all

______ 2 – The author acknowledges coder training, but provides no details

______ 3 – A brief description of coder training is provided

______ 4 – Coder training is described, but additional research is necessary to replicate the
coder training procedure

______ 5 – Coder training is described in enough detail that would be easy to replicate the coder
training procedures

You may add information here to justify your response to Variable 4c if you wish: ________

_________________________________________________________

EXACT DESCRIPTION OF CODER TRAINING (Copy word for word from the
article.)

_________________________________________________________

_________________________________________________________

_________________________________________________________
4d. Does the author describe a specific reliability criterion which establishes consistency (e.g., stopping rule) to determine that coders were adequately trained?

_____ No
_____ Yes [Reliability Criterion (will range from 0.00-1.00): ___________]

**DEFINITION:**
Reliability Criterion: A standardized statistic to indicate reliability (measured on a scale of 0-1).

5. **CODE BOOK**

5a. How does the author describe the code book? (Check one.)

_____ 1 – Code book is kept secret; not described at all
_____ 2 – The author acknowledges use of a code book, but provides no details
_____ 3 – A brief description of the code book is provided
_____ 4 – Code book is described, but additional research is necessary to replicate the study
_____ 5 – Code book is described in enough detail that would be easy to replicate the study

**DEFINITION:**
Code book- Code books provide detailed direction to coders and create a historical record of the coding process. Code books may contain the following information: definitions of primary constructs, descriptions of general coding categories, and an overview of the data entry and management process.

5b. Do the authors state that a copy of the code book (or coding protocol) is available upon request?

_____ Yes
_____ No

5c. Does the text state how coding discrepancies were resolved?

_____ Yes
_____ No

**DEFINITION:**
Coding discrepancies- Instances in which coders do not agree on the coding of a particular variable.
5d. How does the author describe the resolution of coding discrepancies?

_____1 – Method of resolution is kept secret; not described at all
_____2 – The author acknowledges discrepancies were resolved, but provides no details
_____3 – A brief description of the method of resolution is provided
_____4 – Method of resolution is described, but additional research is necessary to replicate the resolution method
_____5 – Method of resolution is described in enough detail that it would be easy to replicate

EXACT DESCRIPTION OF RESOLUTION PROCEDURE (Copy word for word from the article.)

6. ADDRESSING INTERCODER RELIABILITY

6a. Is reliability acknowledged in the article?

_____Yes
_____No

DEFINITION:
Reliability- The extent to which the coding instrument in the study produces the same results when used by different coders; intercoder agreement
NOTES:
The text need not use the words "coders," "reliability," or "agreement" as long as it clearly discusses the attempt to have those involved in analyzing the primary studies (i.e., coders) use the instrument in the same way.

Details about coders, coder training, sampling procedures, and other such methods-related issues are NOT NECESSARILY included in the discussion of reliability. Always apply the rule above.

If the text states that any part of the sample was coded by more than one coder (e.g., "All coders coded all studies in the sample," "The two coders coded 20 of the same studies"), assume that this information IS about reliability.

This variable refers to intercoder reliability; do not code discussion of test-retest reliability, internal reliability, or general information about proper coding procedures as "yes." But remember that ANY mention of intercoder reliability, no matter how brief, counts as discussion of reliability.

Discussion about gathering the sample should be coded as discussion about reliability ONLY IF the text states that two or more people had to agree in order for each unit to be included in the sample (e.g., "Decisions on a primary study’s inclusion in the database were made on a case-by-case basis by three judges").

This includes ANY information that deals with intercoder reliability as described above, not just information about assessing reliability. Therefore, any discussion of attempts to be reliable after a reliability value has been determined should be coded as "yes." This includes discussions of differences in how coders use the instrument and the implications of those differences in terms of the results.

If a text states that coders coded "independently," this information alone is not enough to code as being about reliability.

The text need not report the actual reliability value(s) in order to code this variable as "Yes."

If you are unsure whether a particular sentence is about reliability, and it is between two sentences that are about reliability or it is the first or last sentence in a paragraph about reliability, read carefully -- this sentence is likely to deal with reliability issues as well even if they are not clearly stated (easy to find).

GENERAL NOTE FOR ALL REMAINING VARIABLES:
When coding the remaining variables, it is often necessary to make assumptions. However, be guided by the following rule: If you are certain about how the study did something (e.g., number of coders, nature of reliability dataset, overall sample size), but it is not Clearly stated (easy to find) in the text, code based on your assumption of how they did it and note where appropriate that the information was not clearly stated (easy to find).

6b. How did the study author(s) present reliability information in the text?

______Separate section with "reliability" in heading in text
______One or more complete paragraphs about reliability in text
______Less than a paragraph about reliability in text
______Reliability discussed in FOOTNOTE(S) ONLY
DEFINITIONS:
Separate section with "reliability" in heading in text- one or more paragraphs presented under a subheading that must contain the word "reliability." All information provided under this subheading need not be specifically about reliability.

Complete Paragraph(s) about reliability in text- one or more paragraphs (of two or more sentences each) are devoted to discussing reliability alone, but do not have a subheading containing the word "reliability." The paragraphs can be separated by other text. All paragraphs do not have to be complete, e.g., this includes cases in which there is one complete paragraph and one sentence.

Less than a paragraph about reliability in text - reliability is mentioned either in a paragraph with other information, in several different paragraphs with other information, or in a single sentence that is separated from other text.

Reliability discussed in FOOTNOTES ONLY - there is no mention of reliability in the text, but such information IS discussed in one or more footnotes.

NOTES:
See notes under variable 6a in order to determine what constitutes information about reliability.

If it is difficult to decide whether a particular sentence should be included under this variable, consider whether the sentence alone, without any other discussion of reliability, would constitute reliability being discussed (coded as "yes" under variable 6a).

Supplementary information that refers to a previous discussion of reliability in the text, but is not exactly about reliability (e.g., the sentence "This was considered to be acceptable" after a sentence about reliability) should be included in the number of sentences about reliability.

Remember that reliability information may not be located in one section; it can be discussed in different areas throughout the text.

6c. How many sentences were devoted to the discussion of intercoder reliability (in general) in the text?

_______Number of sentences (You may enter "0.")

NOTES:
Do not include sentences found in footnote(s) in the "number of sentences about reliability in text."

Do not include sentences devoted to the reporting of reliability statistics (e.g., The average intercoder reliability for variables 1-3 was .85.). These will be accounted for in variable 6d.

See notes under variable 6a in order to determine what constitutes information about reliability.

If it is difficult to decide whether a particular sentence should be included under this variable, consider whether the sentence alone, without any other discussion of reliability, would constitute reliability being discussed (coded as "yes" under variable 6a).

Supplementary information that refers to a previous discussion of reliability in the text, but is not exactly about reliability (e.g., the sentence "This was considered to be acceptable" after a sentence about reliability) should be included in the number of sentences about reliability.
6d. How many sentences were devoted to the reporting of intercoder reliability statistics in the text?

_______Number of sentences (You may enter "0.")

NOTES:
Do not include sentences found in footnote(s) in the "number of sentences about reliability in text."

Do not include sentences used to discuss reliability. These sentences were accounted for in variable 6c.

If it is difficult to decide whether a particular sentence should be included under this variable, consider whether the sentence alone, without any other reporting of reliability results, would constitute disseminating reliability results.

6e. Does the article include at least one citation (reference) about intercoder reliability?

_____Yes
_____No

DEFINITION:
Citation- The author has referenced another author's work. This can be determined by use of APA parenthetical documentation or by using footnotes or endnotes.

NOTE:
Remember that the citation must be cited specifically because it is in reference to a reliability issue in order to be included here.

6f. How many citations were related to intercoder reliability?

_______Number of citations (You may enter "0.")

6g. If the article has citations related to reliability, please list the author(s) last name(s) and year.

Author(s)_________________________ Year__________
Author(s)_________________________ Year__________
Author(s)_________________________ Year__________
Author(s)_________________________ Year__________
Author(s)_________________________ Year__________
Author(s)_________________________ Year__________

If there are more citations, please add paper as necessary and STAPLE to the coding sheet.

NOTE:
Remember that the citation must be cited specifically because it is in reference to a reliability issue in order to be included here.
7. REPORTING INTERCODER RELIABILITY

7a. Does the author report a specific method for calculating intercoder reliability?

____ Yes
____ No

NOTE:
Code “No” if the text uses the word “reliability,” but no other term is used to describe the reliability method.

7b. What method was used to report intercoder reliability? (CHECK ALL THAT APPLY):

<table>
<thead>
<tr>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>______ “simple agreement” only</td>
</tr>
<tr>
<td>______ “percentage agreement” or “agreement rate” only</td>
</tr>
<tr>
<td>______ “intercoder reliability” only</td>
</tr>
<tr>
<td>______ intercoder agreement</td>
</tr>
<tr>
<td>______ Cohen’s kappa (or “kappa”)</td>
</tr>
<tr>
<td>______ weighted kappa</td>
</tr>
<tr>
<td>______ Fleiss’ kappa</td>
</tr>
<tr>
<td>______ Holstí’s C.R.</td>
</tr>
<tr>
<td>______ Scott’s pi</td>
</tr>
<tr>
<td>______ Krippendorff’s alpha</td>
</tr>
<tr>
<td>______ Pearson’s $r$, correlation coefficient, intercoder correlation</td>
</tr>
<tr>
<td>______ intraclass correlation coefficient</td>
</tr>
<tr>
<td>______ generialability coefficient</td>
</tr>
<tr>
<td>______ other: ________________________________________________________</td>
</tr>
<tr>
<td>______ method not reported</td>
</tr>
</tbody>
</table>
DEFINITIONS:
Other- select this option and provide the name of any method that is not listed above, or any
generic term used to refer to the reliability value.

Method not reported- the text uses the word "reliability," but no other term is used to describe the
reliability method.

NOTES:
For the options "simple agreement only," "percentage agreement only," and "intercoder reliability
only," select only if the name of a specific formula (e.g., one of the other options) or a more
specific name for a method (e.g., percentage of exact agreement) is NOT listed. In other words,
code these terms only when there is no other option. Keep in mind that these options end with
"only," so even though this variable is "check all that apply," no other options can be selected if
you choose one of these three options.

In order to select one of the listed methods, the text must name the method specifically, e.g., if
the text says "percentage of agreement," do not code as "percentage agreement" or if the text
adds another term to the name such as "intercoder reliability coefficient" do not code as
"intercoder reliability."
Ignore hyphenation and capitalization (e.g., "Inter-coder reliability" = "intercoder reliability").

DO NOT list generic terms (e.g., agreement) under the "other" category if a specific formula is also listed.

If a text cites the work of someone who has developed a reliability formula, DO NOT assume that the formula was used unless the text specifically states that the formula was used.

7c. How did the author(s) report reliability statistics?

______ The author(s) reported reliability statistics only in a text description.

______ The author(s) reported reliability statistics only in a table(s).

______ The author(s) used a combination of text description and table(s) to report reliability statistics.

______ The author(s) did not report reliability statistics.

DEFINITION:
Reliability statistics- a standardized reliability criterion (measured on a scale of 0-1).

NOTE:
If an author reports reliability in terms of a percentage, that value can be converted into a standardized reliability criterion.

7d. If reliability statistics (in terms of a standardized reliability criterion) are not reported, note any phrase in the text that implies a certain level of reliability or coder agreement:

________________________________________________________________________

________________________________________________________________________

________________________________________________________________________

________________________________________________________________________

______ No reliability statistics are reported and no phrase(s) in the text imply a certain level of reliability or coder agreement.

DEFINITION:
Reliability statistics- a standardized reliability criterion (measured on a scale of 0-1).

NOTES:
A common example of such a phrase is "generally agreed."
[For variables 7e – 7u, use all relevant information to code, it may be found in the text, tables, and/or footnotes. Variables 7e – 7u refer to standardized reliability criteria which are reported on a scale of 0-1.]

7e. Did the author differentiate between effect size coding and study descriptor coding?

_____ Yes
_____ No (Skip to Variable 7p.)

DEFINITIONS:
Effect size coding- Coding related to the effect size statistics used in the meta-analysis
Study descriptor coding- Coding related to the study descriptors used in the meta-analysis

7f. Is an overall/average reliability reported for a group or groups of variables (two or more) related to effect size in the study?

_____ Yes
_____ No
_____ Standardized reliability criterion values not reported

NOTES:
If the study reports a reliability figure but does not indicate that it is the highest, lowest, or the reliability for a specific variable, assume that it is an overall/average reliability.

If it is not clear in the text whether a reliability criterion refers to a set of variables or to a single variable that has several categories, or if it is a "check all that apply" variable having several categories for which to respond yes or no, code it as a single (one) variable.

If reliability is reported for an instrument or instruments used in the study, then select "yes" for this variable (i.e., assume that the reliability is an overall/average reliability for a group or groups of variables).

7g. Is a range of reliability reported for a group or groups of variables (two or more) related to effect size in the study?

_____ Yes
_____ No
_____ Standardized reliability criterion values not reported

NOTE:
Examples of how ranges are reported include: "All variables had a reliability figure of .90 or higher," "Reliabilities ranged from .81 to .94," "All variables had a reliability of at least .80."
7h. Is the specific reliability for one or more variables related to effect size reported?

______Yes
______No
______Standardized reliability criterion values not reported

NOTE: If it is not clear in the text whether a reliability criterion refers to a set of variables or to a single variable that has several options, assume that it refers to the single (one) variable.

7i. Are all variables in the study related to effect size accounted for in the reporting of reliability (i.e., is reliability for all variables reported in one of the above three ways in the text)?

______Yes
______No
______Standardized reliability criterion values not reported

NOTES:
To code as "yes," reliability for all dependent variables must be reported.

If the text names a variable other than the ones for which a reliability value is reported, then do not select "yes" (even if results are not reported for that variable). There is one exception to this rule: reliability need not be reported for variables that are treated only as qualitative variables in the study (i.e., the study provides no quantitative measures for these variables).

Select "yes" if reliability for all variables in the study was reported regardless of how few variables are included in the study.

If the text provides an overall/average reliability, assume that it is for all variables unless otherwise indicated. In such cases in which the text does not clearly state that a reliability refers to all variables, code 7i as "no."

7j. Was the information needed to code variables 7e – 7i (effect size variables) clearly stated (easy to find) in the text?

______Yes, the information was clearly stated (easy to find).
______No, I had to make assumptions in order to code the variables.
______Standardized reliability criterion values not reported

NOTES:
In order to code as "yes," the text must use clear language to express the type of standardized reliability reported. If there is any confusion as to what type of reliability is being reported or whether or not the reliability reported refers to all variables, code as "no."
Often the factor that makes variables 7e – 7h unclear is difficulty in determining all variables included in the study of interest that makes it difficult to determine if all variables are accounted for in the reporting of reliability. If it is necessary to assume that all variables are accounted for (or not accounted for) because of such difficulty, code 7i as "No."

7k. Is an overall/average reliability reported for a group or groups of variables (two or more) related to study descriptors in the study?

______Yes
______No
______Standardized reliability criterion values not reported

NOTES:
If the study reports a reliability figure but does not indicate that it is the highest, lowest, or the reliability for a specific variable, assume that it is an overall/average reliability.

If it is not clear in the text whether a reliability criterion refers to a set of variables or to a single variable that has several categories, or if it is a "check all that apply" variable having several categories for which to respond yes or no, code it as a single (one) variable.

If reliability is reported for an instrument or instruments used in the study, then select "yes" for this variable (i.e., assume that the reliability is an overall/average reliability for a group or groups of variables).

7l. Is a range of reliability reported for a group or groups of variables (two or more) related to study descriptors in the study?

______Yes
______No
______Standardized reliability criterion values not reported

NOTE:
Examples of how ranges are reported include: "All variables had a reliability figure of .90 or higher," "Reliabilities ranged from .81 to .94," "All variables had a reliability of at least .80."

7m. Is the specific reliability for one or more variables related to study descriptors reported?

______Yes
______No
______Standardized reliability criterion values not reported

NOTE:
If it is not clear in the text whether a reliability criterion refers to a set of variables or to a single variable that has several options, assume that it refers to the single (one) variable.
7n. Are all variables in the study related to *study descriptors* accounted for in the reporting of reliability (i.e., is reliability for all variables reported in one of the above three ways in the text)?

_____ Yes

_____ No

_____ Standardized reliability criterion values not reported

**NOTES:**
To code as "yes," reliability for all dependent variables must be reported.

If the text names a variable other than the ones for which a reliability value is reported, then do not select "yes" (even if results are not reported for that variable). There is one exception to this rule: reliability need not be reported for variables that are treated only as qualitative variables in the study (i.e., the study provides no quantitative measures for these variables).

Select "yes" if reliability for all variables in the study was reported regardless of how few variables are included in the study.

If the text provides an overall/average reliability, assume that it is for all variables unless otherwise indicated. In such cases in which the text does not clearly state that a reliability refers to all variables, code 7i as "no."

7o. Was the information needed to code variables 7k – 7n (*study descriptor variables*) clearly stated (easy to find) in the text?

_____ Yes, the information was clearly stated (easy to find).

_____ No, I had to make assumptions in order to code the variables.

_____ Standardized reliability criterion values not reported

**NOTES:**
In order to code as "yes," the text must use clear language to express the type of standardized reliability reported. If there is any confusion as to what type of reliability is being reported or whether or not the reliability reported refers to all variables, code as "no."

Often the factor that makes variables 7k – 7n unclear is difficulty in determining all variables included in the study of interest that makes it difficult to determine if all variables are accounted for in the reporting of reliability. If it is necessary to assume that all variables are accounted for (or not accounted for) because of such difficulty, code 7o as "No."
If you completed variables 7f – 7o, skip to variable 7u.

7p. Is an overall/average reliability reported for a group or groups of variables (two or more) in the study?

_____Yes

_____No

_____Standardized reliability criterion values not reported

NOTES:
If the study reports a reliability figure but does not indicate that it is the highest, lowest, or the reliability for a specific variable, assume that it is an overall/average reliability.

If it is not clear in the text whether a reliability criterion refers to a set of variables or to a single variable that has several categories, or if it is a "check all that apply" variable having several categories for which to respond yes or no, code it as a single (one) variable.

If reliability is reported for an instrument or instruments used in the study, then select "yes" for this item (i.e., assume that the reliability is an overall/average reliability for a group or groups of variables).

7q. Is a range of reliability reported for a group or groups of variables (two or more) in the study?

_____Yes

_____No

_____Standardized reliability criterion values not reported

NOTE:
Examples of how ranges are reported include: "All variables had a reliability figure of .90 or higher," "Reliabilities ranged from .81 to .94," "All items had a reliability of at least .80."

7r. Is the specific reliability for one or more variables reported?

_____Yes

_____No

_____Standardized reliability criterion values not reported

NOTE:
If it is not clear in the text whether a reliability criterion refers to a set of variables or to a single variable that has several options, assume that it refers to the single (one) variable.
7s. Are all variables in the study accounted for in the reporting of reliability (i.e., is reliability for all variables reported in one of the above three ways in the text)?

[ ] Yes
[ ] No
[ ] Standardized reliability criterion values not reported

**NOTES:**
To code as "yes," reliability for all dependent variables must be reported.

If the text names a variable other than the ones for which a reliability value is reported, then do not select "yes" (even if results are not reported for that variable). There is one exception to this rule: reliability need not be reported for variables that are treated only as qualitative variables in the study (i.e., the study provides no quantitative measures for these variables).

Select "yes" if reliability for all variables in the study was reported regardless of how few variables are included in the study.

If the text provides an overall/average reliability, assume that it is for all variables unless otherwise indicated. In such cases in which the text does not clearly state that a reliability refers to all variables, code 7t as "no."

7t. Was the information needed to code items 7p – 7s clearly stated (easy to find) in the text?

[ ] Yes, the information was clearly stated (easy to find)
[ ] No, I had to make assumptions in order to code the variables
[ ] Standardized reliability not reported

**NOTES:**
In order to code as "yes," the text must use clear language to express the type of standardized reliability reported. If there is any confusion as to what type of reliability is being reported or whether or not the reliability reported refers to all variables, code as "no."

Often the factor that makes variables 7p – 7s unclear is difficulty in determining all variables included in the study of interest that makes it difficult to determine if all variables are accounted for in the reporting of reliability. If it is necessary to assume that all variables are accounted for (or not accounted for) because of such difficulty, code 7t as "No."

7u. Did the author(s) include a discussion related to the “acceptability” of reliability criterion/values reported?

[ ] Yes
[ ] No

**NOTE:**
In order to check "Yes," the author must have provided a narrative justifying why the obtained reliability values were or were not acceptable.
7v.

What was the LOWEST accepted reliability criterion reported? If MORE THAN ONE reliability method was reported in Variable 7b, list the lowest accepted reliability criterion for each method.

________ Standardized reliability not reported

Method: ________________ Reliability Criterion (will range from 0.00-1.00): __________

Method: ________________ Reliability Criterion (will range from 0.00-1.00): __________

Method: ________________ Reliability Criterion (will range from 0.00-1.00): __________

Method: ________________ Reliability Criterion (will range from 0.00-1.00): __________

Method: ________________ Reliability Criterion (will range from 0.00-1.00): __________

What was the HIGHEST accepted reliability criterion reported? If MORE THAN ONE reliability method was reported in Variable 7b, list the lowest accepted reliability criterion for each method.

________ Standardized reliability not reported

Method: ________________ Reliability Criterion (will range from 0.00-1.00): __________

Method: ________________ Reliability Criterion (will range from 0.00-1.00): __________

Method: ________________ Reliability Criterion (will range from 0.00-1.00): __________

Method: ________________ Reliability Criterion (will range from 0.00-1.00): __________

Method: ________________ Reliability Criterion (will range from 0.00-1.00): __________

DEFINITION:
Accepted reliability criterion- Variable must be used in the meta-analysis for the reliability to be accepted. That is, if a variable is removed from the study because of a low reliability criterion, then it is NOT an accepted criterion.

7w. Was the information needed to code variables 7v clearly stated (easy to find) in the text?

________ No, I had to make assumptions in order to code these variables.

________ Yes, the information was clearly stated (easy to find).

________ Standardized reliability not reported

NOTE:
In order to code as "yes," the text must use clear language to express the type of standardized reliability reported. If there is any confusion as to what type of reliability is being reported or whether or not the reliability reported refers to all variables, code as "no."
7x. What was the lowest reliability criterion deemed “acceptable” by the author?

Lowest Acceptable Reliability Criterion (will range from 0.00-1.00): _____________

________ Standardized reliability criterion values not reported

**DEFINITION:**
Accepted reliability criterion- Variable must be used in the meta-analysis for the reliability to be accepted. That is, if a variable is removed from the study because of a low reliability criterion, then it is NOT an accepted criterion.

**NOTE:**
This value should appear in the lowest accepted reliability criterions table that was created in variable 7j.
Appendix B

Assessment and Reporting of Intercoder Reliability in Published Meta-Analyses

Coding Sheet

Spring 2006

Coder: ___________________________       Date Coded: ___________________________

Article Number: ___________________       Date Entered: _______________________

1. What is the content area of the meta-analysis? (Please rank your decisions with “1” indicating the most appropriate choice(s). You may choose multiple content areas if applicable. You may assign a decision rank to more than one category, i.e., you may indicate that two content areas are the most appropriate choice by assigning them BOTH a rank of “1”.)

   _____ Academic (Educational)       _____ Language Arts Education
       Achievement
   _____ Arts Education                _____ Mathematics Education
   _____ Computer-Assisted Instruction _____ Middle & High School Leadership
   _____ Consumer Sciences             _____ Physical Education
   _____ English as a Second Language  _____ Science Education
   _____ Elementary School Leadership  _____ Social Studies Education
   _____ Foreign Language Education    _____ Special Education
   _____ General Early Childhood Education  _____ School Psychology/Guidance Counseling
   _____ General Elementary Education  _____ Technical Education
   _____ General Middle & High School Education  _____ Vocational Education
   _____ Health Education             _____ Other: ___________________________

2. How many primary studies are included in the meta-analysis? ______
3. CODERS and CODING STRATEGY

3a. Did the author report the number of coders involved in coding the primary studies?
   _____ Yes
   _____ No

3b. How many coders were involved in coding the primary studies?
   (Check all that apply.)
   _____ One coder
   _____ More than one coder (List specific number of coders: ______________, if number of coders is not reported, enter NR.)
   _____ The author does not report any details about the person(s) involved in the coding process.

3c. Was the information needed to code variable 3b clearly stated (easy to find) in the text?
   _____ Yes, the information was clearly stated (easy to find).
   _____ No, the information was not clear. I had to make assumptions to code this variable.
   _____ I was unable to ascertain any information about the individuals involved in coding the primary studies.

3d. Did the author(s) differentiate between coding of the reliability sample and coding of the actual sample?
   _____ Yes
   _____ No

3e. What was the size of the reliability sample?
   _____ Number of primary studies in the reliability sample (You may enter “0.” You may enter the overall sample size if appropriate.)
   _____ The authors did not report any information regarding the assessment of reliability.

3f. Was the information needed to code variable 3e clearly stated (easy to find) in the text?
   _____ Yes, the information was clearly stated (easy to find).
   _____ No, the information was not clear. I had to make assumptions to code this variable.
   _____ The authors did not report any information regarding the assessment of reliability.
3g. What proportion of the primary studies were included in the reliability sample?

_____Total number of primary studies coded (See variable 2.)

_____Total number of primary studies included in the reliability sample (See variable 3e.)

_____Proportion of primary studies coded by multiple coders (Total number of studies coded by multiple coders divided by total number of studies coded.)

_____The author does not present enough information in the text to code this variable.

3h. Did the author provide a description of all of the coders’ levels of expertise?

_____Yes

_____No

3i. How did the author describe the coders’ levels of expertise?

_____Graduate students/graduate research assistants

_____Undergraduate students/undergraduate research assistants

_____Study authors

_____Other: ________________________________________________

4. CODER TRAINING

4a. Does the text state how coders were trained?

_____Yes

_____No

4b. Amount of training (fill in amount; then check unit):

Number: __________ of

_____Hours

_____Units

_____Sessions

_____OTHER__________________________________________________

_____AMOUNT OF TRAINING NOT REPORTED
4c. How does the author describe the coder training procedure? (Check one.)

_____ 1 – Coder training is kept secret; not described at all

_____ 2 – The author acknowledges coder training, but provides no details

_____ 3 – A brief description of coder training is provided

_____ 4 – Coder training is described, but additional research is necessary to replicate coder training procedures

_____ 5 – Coder training is described in enough detail that would be easy to replicate coder training procedures

You may add information here to justify your response to Variable 4c if you wish: __________

__________________________________________________________

__________________________________________________________

__________________________________________________________

EXACT DESCRIPTION OF CODER TRAINING (Copy word for word from the article.)

__________________________________________________________

__________________________________________________________

__________________________________________________________

4d. Does the author describe a specific reliability criterion which establishes consistency (e.g., stopping rule) to determine that coders were adequately trained?

_____ No

_____ Yes [Reliability Criterion (will range from 0.00-1.00): __________]

5. CODE BOOK

5a. How does the author describe the code book? (Check one.)

_____ 1 – Code book is kept secret; not described at all

_____ 2 – The author acknowledges use of a code book, but provides no details

_____ 3 – A brief description of the code book is provided

_____ 4 – Code book is described, but additional research is necessary to replicate the study

_____ 5 – Code book is described in enough detail so that would be easy to replicate the study
5b. Do the authors state that a copy of the code book (or coding protocol) is available upon request?

_____ Yes
_____ No

5c. Does the text state how coding discrepancies were resolved?

_____ Yes
_____ No

5d. How does the author describe the resolution of coding discrepancies?

_____ 1 – Method of resolution is kept secret; not described at all
_____ 2 – The author acknowledges discrepancies were resolved, but provides no details
_____ 3 – A brief description of the method of resolution is provided
_____ 4 – Method of resolution is described, but additional research is necessary to replicate the resolution method
_____ 5 – Method of resolution is described in enough detail that it would be easy to replicate

EXACT DESCRIPTION OF RESOLUTION PROCEDURE (Copy word for word from the article.)

_________________________________________________________________________
_________________________________________________________________________
_________________________________________________________________________
_________________________________________________________________________

6. ADDRESSING INTERCODER RELIABILITY

6a. Is reliability acknowledged in the article?

_____ Yes
_____ No
6b. How did the study author(s) present reliability information in the text?

- Separate section with "reliability" in heading in text
- One or more complete paragraphs about reliability in text
- Less than a paragraph about reliability in text
- Reliability discussed in FOOTNOTE(S) ONLY

6c. How many sentences were devoted to the discussion of intercoder reliability (in general) in the text?

- Number of sentences (You may enter "0.")

6d. How many sentences were devoted to the reporting of intercoder reliability statistics in the text?

- Number of sentences (You may enter "0.")

6e. Does the article include at least one citation (reference) about intercoder reliability?

- Yes
- No

6f. How many citations were related to intercoder reliability?

- Number of citations (You may enter "0.")

6g. If the article has citations related to reliability, please list the author(s) last name(s) and year.

<table>
<thead>
<tr>
<th>Author(s)</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

If there are more citations, please add paper as necessary and STAPLE to the coding sheet.
7. REPORTING INTERCODER RELIABILITY

7a. Does the author report a specific method for calculating intercoder reliability?

_____ Yes
_____ No

NOTE:
Code “No” if the text uses the word “reliability,” but no other term is used to describe the reliability method.

7b. What method was used to report intercoder reliability? (CHECK ALL THAT APPLY):

_____ “simple agreement” only
_____ “percentage agreement” or “agreement rate” only
_____ “intercoder reliability” only

_____ intercoder agreement
_____ Cohen’s kappa (or “kappa”)
_____ weighted kappa
_____ Fleiss’ kappa
_____ Holsti’s C.R.
_____ Scott’s pi
_____ Krippendorff’s alpha

_____ Pearson’s \( r \), correlation coefficient, intercoder correlation
_____ intraclass correlation coefficient
_____ generalizability coefficient

_____ other: ______________________

_____ method not reported
7c. How did the author(s) report reliability statistics?

_______ The author(s) reported reliability statistics only in a text description.

_______ The author(s) reported reliability statistics only in a table(s).

_______ The author(s) used a combination of text description and table(s) to report reliability statistics.

_______ The author(s) did not report reliability statistics.

7d. If reliability statistics (in terms of a standardized reliability criterion) are not reported, note any phrase in the text that implies a certain level of reliability or coder agreement:

______________________________________________________________________________________

______________________________________________________________________________________

_______ No reliability statistics are reported and no phrase(s) in the text imply a certain level of reliability or coder agreement.
[For variables 7e – 7u, use all relevant information to code, it may be found in the text, tables, and/or footnotes. Variables 7e – 7u refer to standardized reliability criteria which are reported on a scale of 0-1.]

7e. Did the author differentiate between effect size coding and study descriptor coding?
   
   _____ Yes
   
   _____ No (Skip to Variable 7p.)

7f. Is an overall/average reliability reported for a group or groups of variables (two or more) related to effect size in the study?
   
   _____ Yes
   
   _____ No
   
   _____ Standardized reliability criterion values not reported

7g. Is a range of reliability reported for a group or groups of variables (two or more) related to effect size in the study?
   
   _____ Yes
   
   _____ No
   
   _____ Standardized reliability criterion values not reported

7h. Is the specific reliability for one or more variables related to effect size reported?
   
   _____ Yes
   
   _____ No
   
   _____ Standardized reliability criterion values not reported

7i. Are all variables in the study related to effect size accounted for in the reporting of reliability (i.e., is reliability for all variables reported in one of the above three ways in the text)?
   
   _____ Yes
   
   _____ No
   
   _____ Standardized reliability criterion values not reported
7j. Was the information needed to code variables 7e – 7i (effect size variables) clearly stated (easy to find) in the text?

- Yes, the information was clearly stated (easy to find).
- No, I had to make assumptions in order to code the variables.
- Standardized reliability criterion values not reported

7k. Is an overall/average reliability reported for a group or groups of variables (two or more) related to study descriptors in the study?

- Yes
- No
- Standardized reliability criterion values not reported

7l. Is a range of reliability reported for a group or groups of variables (two or more) related to study descriptors in the study?

- Yes
- No
- Standardized reliability criterion values not reported

7m. Is the specific reliability for one or more variables related to study descriptors reported?

- Yes
- No
- Standardized reliability criterion values not reported

7n. Are all variables in the study related to study descriptors accounted for in the reporting of reliability (i.e., is reliability for all variables reported in one of the above three ways in the text)?

- Yes
- No
- Standardized reliability criterion values not reported
7o. Was the information needed to code variables 7k – 7n (study descriptor variables) clearly stated (easy to find) in the text?

_____ Yes, the information was clearly stated (easy to find).

_____ No, I had to make assumptions in order to code the variables.

_____ Standardized reliability criterion values not reported

If you completed variables 7f – 7o, skip to variable 7u.

7p. Is an overall/average reliability reported for a group or groups of variables (two or more) in the study?

_____ Yes

_____ No

_____ Standardized reliability criterion values not reported

7q. Is a range of reliability reported for a group or groups of variables (two or more) in the study?

_____ Yes

_____ No

_____ Standardized reliability criterion values not reported

7r. Is the specific reliability for one or more variables reported?

_____ Yes

_____ No

_____ Standardized reliability criterion values not reported

7s. Are all variables in the study accounted for in the reporting of reliability (i.e., is reliability for all variables reported in one of the above three ways in the text)?

_____ Yes

_____ No

_____ Standardized reliability criterion values not reported
7t. Was the information needed to code items 7p – 7s clearly stated (easy to find) in the text?

- Yes, the information was clearly stated (easy to find)
- No, I had to make assumptions in order to code the variables
- Standardized reliability criterion values not reported

7u. Did the author(s) include a discussion related to the “acceptability” of reliability criterion/values reported?

- Yes
- No

7v.

What was the LOWEST accepted reliability criterion reported? If MORE THAN ONE reliability method was reported in Variable 7b, list the lowest accepted reliability criterion for each method.

- Standardized reliability not reported

Method: Reliability Criterion (will range from 0.00-1.00):

Method: Reliability Criterion (will range from 0.00-1.00):

Method: Reliability Criterion (will range from 0.00-1.00):

Method: Reliability Criterion (will range from 0.00-1.00):

Method: Reliability Criterion (will range from 0.00-1.00):

What was the HIGHEST accepted reliability criterion reported? If MORE THAN ONE reliability method was reported in Variable 7b, list the lowest accepted reliability criterion for each method.

- Standardized reliability not reported

Method: Reliability Criterion (will range from 0.00-1.00):

Method: Reliability Criterion (will range from 0.00-1.00):

Method: Reliability Criterion (will range from 0.00-1.00):

Method: Reliability Criterion (will range from 0.00-1.00):

Method: Reliability Criterion (will range from 0.00-1.00):
7w. Was the information needed to code variables 7v clearly stated (easy to find) in the text?

_____ No, I had to make assumptions in order to code these variables.

_____ Yes, the information was clearly stated (easy to find).

_____ Standardized reliability not reported

7x. What was the lowest reliability criterion deemed “acceptable” by the author?

Lowest Acceptable Reliability Criterion (will range from 0.00-1.00): __________

_____ Standardized reliability criterion values not reported
Appendix C


1. Use the following link to access the SPSS website:

http://support.spss.com/tech/Products/SPSS/Documentation/Statistics/articles/whichicc.htm

2. The website will prompt you for a username and password. You can log in to the website as a guest using the following username and password (note that the username and password are not case sensitive):

   Username = guest

   Password = guest