University of Northern Colorado
Scholarship & Creative Works @ Digital UNC

Dissertations Student Work

12-2017

Estimating Bias in Multilvel Reliability Coefficients: A Monte Carlo Simulation

Karen Traxler
University of Northern Colorado

Follow this and additional works at: https://digscholarship.unco.edu/dissertations

Recommended Citation
Traxler, Karen, "Estimating Bias in Multilvel Reliability Coefficients: A Monte Carlo Simulation" (2017). *Dissertations*. 467.
https://digscholarship.unco.edu/dissertations/467

This Dissertation is brought to you for free and open access by the Student Work at Scholarship & Creative Works @ Digital UNC. It has been accepted for inclusion in Dissertations by an authorized administrator of Scholarship & Creative Works @ Digital UNC. For more information, please contact Nicole.Webber@unco.edu.

© 2017

KAREN TRAXLER

ALL RIGHTS RESERVED

UNIVERSITY OF NORTHERN COLORADO

Greeley, Colorado

The Graduate School

ESTIMATING BIAS IN MULTILVEL RELIABILITY COEFFICIENTS: A MONTE CARLO SIMULATION

A Dissertation Submitted in Partial Fulfillment
of the Requirements of the Degree of
Doctor of Philosophy

Karen Traxler

College of Education and Behavioral Sciences
School of Applied Statistics
and Research Methods

December, 2017

This Dissertation by: Karen Traxler

Entitled: *Estimating Bias in Multilevel Reliability Coefficients: A Monte Carlo Simulation*

has been approved as meeting the requirement for the Degree of Doctor of Degree in
College of Education and Behavioral Sciences in School of Applied Statistics and Research Methods

Accepted by the Doctoral Committee

_____
Susan R. Hutchinson, Ph.D., Research Advisor

_____
Trent Lalonde, Ph.D., Committee Member

_____
James Kole, Ph.D., Committee Member

_____
Joyce Weil, Ph.D., Faculty Representative

Date of Dissertation Defense October 23, 2017 .

Accepted by the Graduate School

_____
Linda L. Black, Ed.D.
Associate Provost and Dean
Graduate School and International Admissions

ABSTRACT

Traxler, Karen. *Estimating Bias in Multilevel Reliability Coefficients: A Monte Carlo Simulation*. Published Doctor of Philosophy dissertation, University of Northern Colorado, 2017.

Purpose: The purpose of this dissertation was to generate observed scores under complex data conditions often found in the real world and (a) investigate error in terms of internal consistency reliability within the Classical Test Theory framework (Cronbach's α and polychoric ordinal α) and person reliability within Rasch Rating Scale Model (RSM); (b) inform applied researchers about possible relative bias in reliability coefficients when more complex data structures and underlying distributions are encountered; and (c) provide applied researchers a reference from which to interpret their results. *Methods*: Using Monte Carlo simulation techniques to generate polytomous response choices in single-level and multilevel models, sample reliability coefficients, standard errors of reliability estimates, and levels of absolute relative bias were examined and compared across a range of data conditions, including normal, mixed, and non-normal distributions and varying sample sizes. *Results*: The results support taking the structure of the data collected into account during the analytic phase and provide empirical evidence that if data collected for research are dependent on a higher order structure, reliability coefficients in a multilevel model are less biased than those derived from a single-level model.

Additionally, results support the idea that polychoric ordinal α at level-1 of a two-level sampling design have slightly less bias across all data conditions than Cronbach's α, and under normal and mixed data distributions for person reliability; however, the small gain in the precision of reliability estimates may not be worth the additional effort of calculating polychoric ordinal α for many clinicians and educators. *Recommendations for Applied Researchers*: Using Cronbach's α under normal and mixed data conditions and across sample sizes is acceptable and easier to estimate due to its availability in social science software. For extremely non-normal data, the Rasch-RSM model should be used since the effort is worth the lower level of bias. The results also show that a variety of different data properties jointly affect reliability coefficients and care should be taken to provide both context and a theoretical framework in which to interpret results.

Keywords: *Reliability, Cronbach's α, polychoric ordinal α, multilevel models, multilevel confirmatory factor analysis, Rasch item response theory, rating scale model*

ACKNOWLEDGEMENTS

This dissertation is dedicated to my husband, Joe Harpring, who introduced me to the world of statistics, taught me that nothing is unknowable, encouraged me to return to college and earn my Bachelor's degree, stood by me through my Master's degree, inspired me to continue my education and explore the many facets of behavioral, educational, and social research, and who is certainly ready for me to earn my Ph.D. I must also acknowledge my two sons, Joey and Jesse, who grew up watching their mother spend long hours studying and who learned first hand the value of education and the need for persistence. I also acknowledge my late husband, Mark Traxler, who believed in education and spent long hours working toward his degree. Mark inspired me to continue my education after his death.

Actively pursuing any Ph.D. involves the entire family, friends, and advisors, and they were all my cheerleaders. Through countless hours, late nights, and long days, my family and friends always provided their support. I acknowledge my late-mother-in-law Sherry Harpring who cared for the boys when I had deadlines to meet, and always supported my efforts, my sister-in-law Cindy Harpring, and my father-in-law Ken Harpring for always encouraging me and listening to all my woes, and my good friend Patti Bennett, who not only cared for my late-grandmother-in-law when I was working, but who took care of me and helped me find the time to complete the doctoral program. She dedicated herself to my family over the years, cared for all of us, grieved with us when we lost, not only my grandmother-in law, but my mother-in-law Sherry over a

3-month span, and far too early, and never waivered in her belief in me. I acknowledge my father and step-mother, Jim and Jean Ashley for their words of encouragement, Niloofar Ramenzani and Maryann Shane, my good friends and colleagues for never allowing me to give up, no matter what, and for illuminating the path for my future as a Research Analyst at the Center for Naval Analyses, and my close friend Sheri Hannah-Ruh for providing me opportunities to grow, learn, and flex my evaluative muscles.

I acknowledge my mother and step-father, Phyllis and Wil Wiederaenders, who always believed in me, even when I had my doubts, and who let me set up shop at their home in North Carolina more than once for my week-long writing sprees. Wil even set up the office with everything I would need to succeed, including internet access, lamps for late-night work, sodas, snacks and good humor! My mom continued to encourage me regardless of my mood, listened calmly to my hopes and fears, and offered her sage advice on how to succeed. Thank you mom and Wil…this is for you!

Writing a dissertation requires self-motivation and personal commitment; however, I could not have completed this without the guidance of the ASRM faculty, staff, and students at the University of Northern Colorado. I acknowledge Dr. David Gilliam, my undergraduate "stats" teacher who saw something in me I did not and encouraged me to press forward and Dr. Lisa Rue, from whom I learned to love program evaluations and randomized control trials and received the best training a student could want: hands-on and front-line. Finally, a very special acknowledgement is reserved for Dr. Susan Hutchinson, my mentor from the first day of graduate school in 2009, to the last in 2017. Thank you, everyone.

TABLE OF CONTENTS

CHAPTER

I. INTRODUCTION 1

Measuring Latent Traits 2
An Overview of Measurement Theories 2
Psychometric Analysis of Scores from Assessment Instruments 3

Reliability Within the Classical Test Theory Framework 5
Reliability Within the Rasch Item Response Theory Framework 6

Problem Statement 6
Rationale for the Study 7
Purpose of the Study 8
Research Questions and Hypotheses 9
Limitations 10
Chapter Summary 11

II. REVIEW OF THE LITERATURE 13

Fundamental Measurement 14
Response Scaling 18

Thurstone Response Scaling 19
Likert Response Scaling 21
Guttman Response Scaling 22

Levels of Measurement 24

Response Scaling 25
Dichotomous vs. Polytomous Response Scaling 25

The Evolution of Measurement Theories 28

Introduction to Measurement Theories 28
Classical Test Theory 29

Assumptions of Classical Test Theory 30
Advantages of Classical Test Theory 32

CHAPTER
II. continued

Disadvantages of Classical Test Theory 34
Estimating reliabilty in Classical Test Theory 36

Two Coefficients to Estimate Internal Consistency 37

Cronbach's coefficient alpha 37
Assumptions of coefficient alpha 39

Consequences of Underestimating Alpha 43

Polychoric ordinal α 44
Advantages of polychoric ordinal α 45

Item Response Theory 46
Rating Scale Model 49
Assumptions of Rasch Item Response Theory and the Rating Scale Model 50
Advantages of the Rasch Item Response Theory and the Rating Scale Model 51
Disadvantages of the Rasch Item Response Theory and the Rating Scale Model 51
Estimating Reliability 52

The Importance of Reporting Reliability Estimates 53
Statistical and Psychometric Properties Affected by Reliability 57

Effect Size 57
Validity 60
P-value 60
Power.. 60
Type II Error 61

Factors Affecting Reliability 61

Single-Level Sampling Design 62

Sample size 62
Classical Test Theory 63
Guidelines for estimating reliability 64

CHAPTER

II. continued

Rasch item response theory 68
Guidelines for estimating reliability 68

Number of Items 70
Number of Response Choices 76
Types of Distribution: Normal vs. Non-Normal Data 82
Multilevel Model 84
Sample Size in a Multilevel Model 86
Number of Items, Response Choices, and Distributions 88

Building a Two-Level Model 88

Using Confirmatory Factor Analysis to Estimate Reliability in a Two- Level Model 90
Multilevel Confirmatory Factor Analysis 91
Assessing Reliability Within the Rasch Item Response Theory Framework 101

Chapter Summary 104

III. METHODS 105

Introduction and Research Questions 105
The Pilot Study 106

Cronbach's Alpha 107
Pilot Study Results 109

Single-level Cronbach's α 109
Single-level Person and item reliability 111

Sampling Designs and Data Conditions for the Full Study 116

Single-Level Sampling Design 117
Two-Level Sampling Design 117

CHAPTER

III. continued

Simulation Procedures and Building the Models....124

Within the Classical Test Theory and Multilevel Confirmatory Factor Analysis Frameworks....124
Generating Single-Level Data Sets....124
Estimating Reliability in a Single-Level Sampling Design....125
Generating Multilevel Data Sets....126
Estimating Reliability in a Two-Level Sampling Design....127
Estimating Reliability in the Two-Level Model....129
Within the Rasch Item Response Theory Framework....129
Estimating Reliability Within the Rasch Item Response Theory Framework....130

The single-level model....130
The multilevel sampling design....131

Final Data Conditions Examined....132
Summary of Final Data Conditions....133
Data Analysis....135

Chapter Summary....136

IV. RESULTS....138

Presentation of Results....138

A Recap of Data Conditions and Reliability Terminology....154
Results by Research Questions and Hypotheses....154
Classical Test Theory Single-Level vs. Rasch Rating Scale Model Single-Level Results....154

Reliability and standard errors....154
Relative bias....158

Tests of Hypotheses for Research Question 1....161
Multilevel Model Bias Across Data Conditions....164
Multilevel Reliability Estimates Across Measurement Frameworks....165

Reliability and standard errors....165
Reliability bias....176

CHAPTER

IV. continued

Tests of Hypotheses for Research Question 2 184
Level-1 Reliability Coefficients as the Dependent Variable 198
Level-2 Reliability Coefficients as the Dependent Variable 203
Level-1 Standard Errors of Reliability Estimates 206
Interactions for Standard Errors at Level-1 206
Interactions for Standard Errors at Level-2 210
Level-1 Percentage of Relative Bias Across Data Conditions.....211
Level-2 Percentage of Relative Bias Across Data Conditions.....212
Direction of Relative Bias Across Data Conditions 212
A Comparison of Single-Level and Level-1 Standard Errors and Bias Across Data Conditions 214

Standard errors 214
Reliability bias 215

Test of Hypotheses for Research Question 3 216
Single-Level and Multilevel Interactions Across Data Conditions 226

Conclusions 227

Single-Level Sampling Designs 227

Normal distributions 228
Mixed distributions 229
Non-normal distributions 230
Conclusions for single-level results 230

Multilevel Sampling Designs 230

Normal distributions 231
Mixed distributions 231
Non-normal distributions 231
Conclusions for multilevel results 231

V. CONCLUSIONS 233

Research Question 1: Single-Level Results and Discussion 234

Expected Results 234
Unexpected Results 234

CHAPTER
V. continued

Research Questions 2 through 4: Multilevel Results and Discussion.....236

Expected Results.....236
Unexpected Results.....237

Implications and Recommendations.....239
Limitations.....246
Recommendations for Future Research.....247

REFERENCES.....250

APPENDICES
A. Generating Multivariate Distributions for Crobach's Alpha.....280

B. Generating Multivariate Distributions for Polychoric Ordinal Alpha.....284

C. Generating Multivariate Distributions for Crobach's Alpha-Polychoric Ordinal Alpha-Multilevel.....289

D. Generating Multivariate Distributions for Person Reliability at the Single Level.....293

E. Generating Multivariate Distributions for Person Reliability at the Multilevel.....296

LIST OF TABLES

Table

1. An Example of a Consistent Guttman Cumulative Scale 23
2. Stevens' Four Levels of Measurement 25
3. Frequency of Reliability Estimation Methods Between 1927 and 2001 56
4. Reported Sample Sizes to Estimate Cronbach's Alpha 64
5. Sheng and Sheng's Three Levels of Distribution Conditions 83
6. Summary of Pilot Study Data Conditions 108
7. Pilot Study: Cronbach's α and Corresponding 95% Confidence Intervals from the Single-Level Sampling Design and Normal Distribution 110
8. Tests of Between-subjects Effects for Bias in Cronbach's Alpha in a Single-Level Model 111
9. Pilot Study: Person and Item Reliability and Separation Indices from a Rating Scale Model Single-Level Sampling Design 112
10. ANOVA Results for Absolute Bias Within the Rating Scale Model 113
11. Summary of Final Data Conditions 114
12. Summary of Data Condition Recommendations for Single-Level Models 119
13. Summary of Data Condition Recommendations for Multilevel Models 122
14. Presentation of Single-Level Results 140
15. Presentation of Multilevel Results 141

Table

16. Comparison of Bias Across Measurement Frameworks, Data Conditions, and Data Structures 143

17. Single level Sample Reliability Coeffcient Results 144

18. Absolute Value and Percentage of Bias Across Type of Reliability, Data Distribution, and Sample Size 147

19. Results of Reliability Estimates and Standard Errors for Tests of Hyptheses (ANOVA's 1 and 2).....150

20. Results of Relative Bias and Percentage/Direction of Relative Bias for Tests of Hyptheses (ANOVA's 3 and 4) 151

21. Single-Level Direction of Relative Bias > 10% Across Distribution and Sample Sizes (Chi-square Results) 162

22. Results of Multilevel Sample Reliability Coeffcients 166

23. Percentage of Bias in Multilevel Models for Both Level-1 and Level-2 Across Types of Reliability, Sample Sizes, and Distributions.....177

24. Results for Factrorial ANOVA for Level-1 Reliability Coefficients Across Sample Size and Distribution 186

25. Assessment of Level-1 Reliability Coefficients When Interaction Plots Showed No Interaction 187

26. Results for Factorial ANOVA for Level-2 Reliability Coefficients Across Sample Size and Distribution 188

27. Level-2 Reliability Coefficients' Across Distribution and Sample Size: Assessment of Simple Main Effects 190

28. Level-1 Standard Errors of Reliability Across Distribution and Sample Sizes 191

29. Level-1 Standard Errors of Reliability Coefficients: Main Effects 193

30. Level-1 Relative Bias ≥ 10% Across Distribution and Sample Sizes 194

31. Count of Level-1 Relative Bias ≥ 10% Across All Data Conditions.....195

Table

32. Level-1 Direction of Relative Bias ≥ 10% Across Distribution and Sample Sizes (Chi-square Results) 198

33. Average Standard Errors of Measurement for Reliability Estimates Across Data Conditions 215

34. A Comparison of Single-Level Bias to Level-1 Bias Across Data Conditions 217

35. Single Level and Level-1 Standard Errors of Reliability Estimates Across Data Condition: Assessment of Main Effects 222

36. Results of Factorial ANOVA Comparing Bias ≥ 10% Between Single Level and Level-1 of a Two-Level Sampling Design 223

37. Single-Level and Level-1 Direction of Relative Bias > 10% Across Distribution and Sample Sizes (Chi-square) 225

38. A Single-Level Model Tool for Applied Researchers 242

39. A Two-Level Model Tool for Applied Researchers 244

LIST OF FIGURES

Figure

1. An example of Thurstone's (1928) Continuum of Summated Group Ability 20

2. Wright Map for a Measure of Hope on a 12-item Survey 28

3. A Hypothetical Path Diagram of a Multilevel Confirmatory Factor Analysis Model of Attitude. 99

4. An Example of a Four-Item Two-Level Polytomous Factor Model 128

5. Single level reliability coefficients across distributions and sample size. 157

6. Single level standard errors of reliability across distributions and sample size 158

7, A Visual Representation of Relative Bias Across Data Distributions and Sample Size 160

8. A Visual Representation of the Direction and Percentage of Average Relative Bias in a Single Level Model Across Types of Reliability, Data Distributions, and Sample Size. 161

9. Average Reliability Coefficients at Level-1 of a Two-Level Model Across Data Conditions 173

10. Average Spearman-Brown Coefficients Across Data Conditions. 174

11. Standard Errors in Level-1 of a Two-Level Model Across Data Conditions. 175

12. Level-2 Standard Errors Across Data Conditions. 176

13. Interaction Effects Where the Marginal Means of Reliability Estimates Based on Level-2 Sample Size are Averaged Across the Type of Data Distribution. 200

Figure

14. A Graphical Representation of Interaction Effects of Level-1 Sample Size Marginal Means Average Across Type of Reliability Coefficient.....201

15. A Graphical Representation of Interaction Effects of Level-2 Sample Sizes Across Types of Reliability. 202

16. A Graphical Representation of the Interaction Between Level-2 Sample Sizes and Data Distributions for Level-2 Reliability Coefficients. 205

17. Interaction Effect on Standard Errors of Reliability Coefficients Between Level-1 Sample Sizes and Type of Distribution. 207

18. A Graphical Representation of the Interaction Effect on Level-1 Standard Errors Of Reliability (SERELI) Between Level-1 Sample Size and Type of Reliability Coefficient (Relitype). 208

19. The Interaction Plot Showing the Interaction Between Level-1 and Level-2 Sample Sizes on Estimates of Standard Errors of Reliability (SERELI). 209

.20. Plot of the Interaction Effects Between Level-2 Sample Size and Type of Reliability (Relitype) for Level-2 Standard Errors.....211

21. Cronbach's α and Polychoric Ordinal α Reliability Coefficients, Along with Their Associated Standard Errors of Measurement Across Data Conditions. 229

CHAPTER I

INTRODUCTION

In the United States the emphasis on evidence-based practices (EBPs) in the fields of behavioral, educational, psychological, and social sciences propels the demand for reliable and accurate results on a variety of self-report and objective assessments. Surveys and assessment instruments are a common method used to measure an assortment of individual and group attributes such as attitudes, beliefs, cognitive competencies, abilities, and performance. In many cases, individual certification or licensure are at stake, therefore, clinicians, teachers, administrators, and other stakeholders must be able to depend on the results observed on the assessment instruments employed (Townsend, Christensen, Kreiter, & ZumBrunnen, 2010). The development and implementation of effective treatments, interventions, and programs across the fields of education, psychology, and the social sciences rely on assessments that consistently measure the traits they were developed to measure. Therefore, it is imperative that the systematic processes by which assessments are developed and administered and data are collected and analyzed be established and practiced (Converse, 2009; Thorndike & Thorndike-Christ, 2010). Consequently, it is critical that, educational and social researchers support these stakeholders through the rigorous examination of the methodological issues involved in consistent and valid measurement of the individual and group traits of interest, from attitudes to aptitude. This dissertation focused on the

reliability of scores related to measures of attitude, specifically polytomously scored items using a multilevel sampling design.

Measuring Latent Traits

Quantification is the objective for many social, psychological, and behavioral science researchers (Converse, 2009; Cronbach, 1951; Likert, 1932; Thorndike & Thorndike-Christ, 2010; Thurstone, 1924). Unlike the concrete measurements used in biology, physics, chemistry, and other natural sciences, measurements in socio-behavioral research are more conceptual, requiring abstract thinking and the formation of theoretical constructs, also known as latent traits or factors (Andrich, 1988; Pedhazur & Schmelkin, 1991). In other words, since most phenomena of interest in socio-behavioral sciences are measured indirectly, that is, inferences are drawn based on various indicators related to the traits being studied, social research utilizing self-report and objective assessment tools is seen as a practical method of data collection and the use of these instruments is now widely accepted. However, the debate of using fundamental measurement processes with implicitly measured traits continues. A deeper understanding of the evolution of measurement theories in the social sciences may illuminate the rationale for the methods employed to measure unobserved phenomena.

An Overview of Measurement Theories

Measurement theories and statistical models used to measure latent traits and assess accurate response scaling have evolved and three distinct measurement theories now dominate the research and application of assessment tools: Classical Test Theory (CTT; Spearman, 1904), Generalizability Theory (Brennan, 1992), and Item Response Theory (IRT; Rasch, 1960). Two important components of these prevailing measurement

theories are the understanding and treatment of measurement error (Brennan, 1992; Cronbach, 1951; Guttman, 1950; Likert, 1932, Lord, 1952; Lord & Novik, 1968; Masters, 1982; L. K. Muthén & Muthén, 2002; Rasch, 1960; Spearman, 1904; Thurstone, 1928). In this dissertation I focused on CTT and Rasch IRT measurement theories. Rasch IRT is a subset of IRT which evolved to address some of the limitations of CTT such as sample dependence, the lack of specific item level information, and the inability to partition variance (J. B. Kline, 2005). Both CTT and Rasch IRT use quantitative methods to measure latent traits by assessing the true relationships between empirical observations (Pedhazur & Schmelkin, 1991). Measurement theories are subsumed within these widely accepted measurement frameworks (CTT; and Rasch IRT). Regardless of the measurement framework embraced, by maximizing the consistency and accuracy of the results and minimizing measurement error through the systematic use of well-established methods, measurement frameworks provide the tools necessary to conceptualize individual and/or group differences. Both CTT and Rasch IRT will be discussed in more detail in Chapter II.

Psychometric Analysis of Scores from Assessment Instruments

As mentioned previously, in socio-behavioral research, CTT and Rasch IRT frameworks are widely used to assess the relationships between observed item responses and unobserved latent traits of interest on assessment instruments using psychometric analysis. The National Council on Measurement in Education (Kolen & Tong, 2009) defined psychometrics as “a field of study concerned with the theory and technique of psychological measurement, assessment, and related activities” (para 1). The field of psychometrics encompasses the objective measurement of attitudes and aptitudes as well

as the development and validation of assessment instruments such as personality tests, questionnaires, tests, and raters' judgments. Psychometric analytic techniques are therefore used to examine bias in the observed scores, response scaling, item to sample size ratio, multilevel data structures, and unobserved latent traits measured. These aspects are multifaceted and require intense scrutiny. During the development phase of any assessment tool, reliability and validity are considered "the two most important fundamental characteristics of any [psychometric] procedure" (Miller, 2004, p. 1). Miller (2004) explained that scores on an assessment instrument can be reliable (representing consistency and reproducibility) without being valid (representing accuracy) but cannot be valid without first being reliable. Reliability coefficients are estimates of true measure variance to observed measure variance and since the reliability of scores impacts validity, the intent of this dissertation was to examine any bias in estimates of reliability across a myriad of data conditions and sampling designs. These data conditions include varying sample sizes and single and two-level data structures. The premise being that both data conditions and sampling designs have the potential to introduce measurement error which may render the interpretation of results suspect (Cronbach, 1951; Guttman, 1950; Likert, 1932, Lord, 1952; Lord & Novick, 1968; Masters, 1982; L. K. Muthén & Muthén, 2002; Rasch, 1960; Spearman, 1904; Thurstone, 1928).

Reliability is not an index of quality but a measure of relative reproducibility and as is well-known, reliability is not a property of the instrument itself but of the scores obtained from a particular sample of examinees by the instrument (American Educational Research Association, 2014). Reliability is sample dependent and predicated on the level of measurement (dichotomous, ordinal or continuous scores), distribution of scores,

number of items and response choices, the nature of the relationship between the variables and the latent trait of interest, and any group differences.

Reliability Within the Classical Test Theory Framework

The most common reliability coefficient in published social science literature is Cronbach's α (Cronbach, 1951). Alpha, which emerged from CTT is a coefficient of internal consistency. Cronbach's α is best suited for continuous data, although it is often used for polytomously (ordered) and dichotomously scored (yes/no, true/false, correct/incorrect) data, which are then treated as continuous. The theoretical value of Cronbach's α falls between 0 and 1 and will increase as the inter-item correlations increase (Cronbach, 1951).

An adaptation of Cronbach's α being revisited in contemporary research is the polychoric ordinal α used for polytomously scored variables such as those found in Likert or Likert-type responses (Bonanomi, Ruscone, & Osmetti, 2013; Gadermann, Gruhn, & Zumbo, 2012; Zumbo, Gadermann, & Zeisser, 2007). The polychoric ordinal α utilizes the polychoric correlation coefficient introduced by Pearson (1900). Polychoric ordinal α is also recommended by Ekström (2009), Ekström (2010); Haldgado-Tello, Chacón-Moscoso, Barbero-García, and Vila-Abad (2008), and Zumbo et al. (2007) to measure ordinal variables such as those obtained from an ordinal response scale.

Finally, CTT-based reliability of observed scores can also be estimated using Confirmatory Factor Analysis (CFA) for single-level models and Multilevel Confirmatory Factor Analysis (MCFA) for multilevel models, where the objective is to test whether the observed scores on an assessment instrument fit a hypothesized measurement model T. A. Brown (2015); Geldhof, Preacher, and Zyphur (2014),

Raudenbush and Bryk (1994, 2002). Though there are other methods for producing CTT-based reliability estimates, these are beyond the scope of the current study and are thus not described.

Reliability Within the Rasch Item Response Theory Framework

Person reliability and person separation as well as item reliability and item separation account for reliability estimates within the Rasch IRT family of models. In other words, reliability in Rasch IRT models varies across person ability levels, and depends specifically on how well the items' difficulty matches a person's ability (Bond & Fox, 2014; Rasch, 1960).

Since one main aspect of this dissertation was to focus on polytomously scored (ordinal) assessment items with the same number of response choices across items, the rating scale model (RSM), an extension of the Rasch IRT model, was examined

Problem Statement

With the national call for behavioral, educational, and social interventions, treatments, and programs based on empirical evidence (i.e.: evidence based practices: EBP's), methodological studies regarding the consistency and accuracy of the scores obtained on measurement instruments used to support these interventions and treatments and programs are necessary. Stakeholders and policy-makers alike count on the results of these studies that utilize assessment tools to allocate resources and expand or dismantle programs. Therefore, it is critical that these decisions are predicated on reliable, accurate, and interpretable results, regardless of the complexities of the research design. As mentioned previously, a thorough review of the literature indicates that reliability is one of the most important characteristics of any psychometric procedure, regardless of the

underlying measurement framework (Allen & Yen, 1979; Choi, Dunlop, Chen, & Kim, 2011; Culligan, 2013; Culpepper, 2013; Dick & Hagerty, 1971; Fitzmaurice, 2002; Gaberson, 1997; Gliner, Morgan, & Harmon, 2001; J. B. Kline, 1999, 2005; Shavelson & Webb, 1991; Thorndike & Thorndike-Christ, 2010). Considering the high-stakes decisions based on assessment results, providing guidance on how best to obtain accurate estimates of reliability of the scores on any measurement instrument across multiple disciplines is paramount. Since reliability is heavily affected by item and respondent attributes of latent distributions, the standard error of measurement for any given latent trait value will also be affected by these item and respondent attributes (Culpepper, 2013). Issues related to reliability estimates within the CTT, and IRT frameworks have been well documented; however; bias in estimating reliability coefficients across these frameworks using polytomous data and examining both standard estimates and polychoric coefficients under realistic data circumstances is uncertain, especially in multilevel sampling designs which are discussed in more detail below.

Rationale for the Study

Charter (2003), Cicchetti (1994), Culpepper (2013), Gadermann et al. (2012), Geldhof et al. (2014), Linacre (2012), Maas and Hox (2005), Nunnally and Bernstein (1994), Wright and Stone (1979), Yurdugul (2008); Zumbo et al. (2007), and others suggest building upon previous research related to accurate reliability estimates in CTT and Rasch IRT by further assessing the appropriate sample sizes and shapes of the latent distributions with respect to ordinal response items.

Key considerations when estimating reliability were the level of measurement for the response scale and the underlying structure of the data. For example, with greater

emphasis on EBPs, educational and social researchers must be able to take into account the effects of ordinal response scales and more complex sampling designs on estimates of reliability. These advanced designs are central to their research and the investigation into the distinct sources of error variation must include variable interactions (Bonito, Ruppel, & Keyton, 2012; Davidson, Cooper, & Bullock, 2010; B. O. Muthén, 1994). Few studies have examined the methodological issues inherent in estimating reliability using multilevel modeling (Gadermann et al., 2012; Geldhof et al., 2014; Huang & Cornell, 2016; Raykov & Penev, 2010; Sheng & Sheng, 2012). While these studies assess multilevel data structures under varying data conditions, none of these researchers specifically examined the consequences of non-normal data on reliability coefficients in multilevel models, nor did they assess polytomous data under the concurrent conditions of non-normality and multilevel data even though these complex data structures are a reality in educational and social science research. Finally, previous researchers examining polychoric ordinal α recommend varying sample sizes and distributional characteristics and measuring corresponding levels of bias to contribute to the methodological literature regarding reliability estimation, providing guidance to clinicians, educators, stakeholders, and applied researchers on the consequences of research design decisions on reliability estimates and inform academic, personal, professional, and policy determinations based on assessment results.

Purpose of the Study

The purpose of this dissertation was to assess four reliability coefficients under real world data conditions and sampling designs within the CTT and IRT frameworks by conducting a Monte Carlo simulation. Three main aspects of this dissertation are (a) to

generate polytomously scored sample data which represent a myriad of population data characteristics known to affect the reliability of scores obtained from a hypothetical assessment tool, (b) to assess reliability estimates and standard errors derived from both single-level and two-level models, and c) to investigate and report any bias found in reliability estimates across these data conditions and sampling designs.

Research Questions and Hypotheses

For this dissertation, using Monte Carlo simulation techniques, sample sizes and distributional characteristics were varied and levels of bias in reliability estimates were assessed, reported, and compared, when applicable, across single-level and two-level data structures. Detailed specifications for the varying data conditions and fixed parameters are found in Chapter III of this dissertation. The research questions answered in this study are:

Q1 In a single-level model, to what degree do data conditions (sample size and distribution of data) affect levels of bias in reliability estimates (a comparison of Cronbach's α, polychoric ordinal α, and person reliability)?

H1 In single-level models, bias in reliability estimates will increase under the conditions of smaller sample sizes and non-normal or mixed distributions and polychoric ordinal α and person reliability will be less biased than Cronbach's α.

Q2 In a multilevel model, to what degree do data conditions (sample size and distribution of data) affect levels of bias in reliability estimates (a comparison of Cronbach's α, polychoric ordinal α, and person reliability in level-1 (within groups) and the Spearman-Brown's prophecy coefficient in level-2 (between groups)?

H2 In multilevel models, bias in reliability estimates in level-1 will increase under the conditions of smaller sample sizes and non-normal or mixed distributions and polychoric ordinal α will be less biased than Cronbach's α and person reliability. Additionally, Spearman-Brown's prophecy coefficient will be underestimated under the conditions of smaller sample size and non-normal or mixed distributions

Q3 Do standard errors and levels of bias in reliability estimates (Cronbach's α, polychoric ordinal α, and person reliability) differ when data are single-level versus when data are at level-1 of a two-level across sample size and distribution of data?

H3 When comparing the standard errors and levels of bias in reliability estimates of single-level and level-1 of two-level sampling designs, across three estimates of reliability, bias for level-1 of the two-level model will be lower than the bias found in the single-level models.

Q4 To what degree do interactions among sample size, data distribution, and sampling design (e.g., single-level and two-level) affect levels of bias in reliability estimates (Cronbach's α, polychoric ordinal α, person reliability, and Spearman-Brown's prophecy coefficient)?

H4 Interactions among sample size, data distribution, and sampling design will increase bias in reliability estimates, with the joint effect of lower sample sizes and non-normal and/or mixed distributions displaying the most bias.

Limitations

There are several limitations to this dissertation. First, limitations inherent to Monte Carlo simulation studies include the inability to define or apply context (e.g., theoretical foundations) to the results beyond hypothetical situations. In other words, Monte Carlo simulation procedures are data-intensive experimental designs requiring researchers to make numerous decisions regarding data conditions and sampling designs not always found in real-world data conditions, such as levels of non-normality and varying response patterns. Second, while the ability to control all data conditions selected for the study is alluring, these decisions may result in significant consequences. For example, in this dissertation, I held the number of items and the number of response choices constant for manageability of the design (items = 10, response choices = 5), I fixed Cronbach's α, polychoric ordinal α, and person reliability to .70 and Intraclass

Correlation Coefficients (ICCs) in the multilevel model to .20, and then standardized person ability and item difficulty in order to resemble a fairly well-developed assessment tool administered to an ideal target population. Additionally, I selected only three levels of sample size and three item distributional characteristics (normally distributed data, a mixed data distribution with ½ of the responses normally distributed and ½ of the item responses non-normal, and a fully non-normal distribution) in the single-level models and two levels of sample size, two levels of group size, and three item distributional characteristics in the multilevel model with the intention of replicating real world data conditions. Each of these decisions has consequences on the level of bias in the reliability coefficients. Third, Monte-Carlo simulation will never capture all of the possible data conditions, sampling designs, and crossed designs implemented by applied researchers, limiting the application and generalizability of the results.

Chapter Summary

Measurement frameworks such as CTT and Rasch IRT are the most commonly utilized frameworks to develop, validate, and assess individual and group responses. Since the use of assessment tools, specifically measures of attitude using polytomously scored rating scales developed within these frameworks has increased to meet the emphasis on EBPs in the fields of behavioral, educational, psychological, and social sciences, consequences based on assessment results have intensified. Consequently, methodological studies regarding the reliability of responses has become imperative under a mélange of polytomously scored data, a variety data conditions and two-level models. Currently a considerable amount of methodological literature addresses issues relating to reliability estimation as almost an afterthought, as if the debate surrounding

the ramifications of biased estimates were settled long ago. The truth is, for those handful of researchers interested in the behavior of reliability estimates in the more complex sampling designs emerging in the educational and social sciences, the debate has been renewed with vigor. The practical importance of studies designed to address reliability estimate bias under the more realistic data characteristics found in applied educational and behavioral research, such as small sample sizes and data distributions not meeting the assumptions of normality or independence, cannot be overemphasized. Building on contemporary research conducted by Huang and Cornell (2016), Little (2013), Geldhof et al. (2014), Raykov and Penev (2010), and Sheng and Sheng (2012), through this dissertation I endeavored to fill in some of the gaps in the literature regarding bias in reliability estimation and generalization. Chapter II presents the theoretical and research literature supporting the need, purpose, data conditions, and distributional characteristics used in this dissertation, with a thorough examination of the importance of calculating and reporting reliability coefficients and the need to understand the role measurement error plays in estimating reliability. Chapter III provides a detailed description of the methods used to generate data and examine bias in reliability estimates across all data conditions and sampling distributions for single-level and multilevel models. The results are presented in Chapter IV, organized by research question and sampling design, In Chapter V, I communicate my conclusions and recommendations for applied researchers, clinicians, and educators and provide practical guidance on interpreting reliability coefficients under varying data conditions.

CHAPTER II

REVIEW OF THE LITERATURE

This review of the literature provides the empirical basis to warrant, not only the need for the current study, but the specific research questions introduced in Chapter I. To understand the full scope of the myriad of issues related to accurate reliability estimates of scores obtained from summated rating scales (such as those found in psychological and educational research), Chapter II begins with a discussion of fundamental measurement in the realm of psychological and educational assessment. This is proceeded by a reflection on the origins of contemporary scaling methods and their relationship to fundamental measurement. Included in this section are issues related to item response scaling and the development of the Thorndike (1919), Thurstone (1928), Likert (1932), and Guttman (1950) response scaling methods. Next, levels of measurement, recommended by Stevens (1946, 1951) as a useful way to classify variables, are described. Data classification and types of data, such as dichotomous or polytomous, are then explored and an additive conjoint model is introduced.

Following a thorough review of the foundations of response scaling and item calibration, two of the most commonly used frameworks of measurement are presented and defined: CTT (Spearman, 1904) and Rasch IRT (Lord, 1952; Lord & Novick, 1968; Rasch, 1960). Since these measurement frameworks carry a set of assumptions regarding the underlying structure and distribution of data and contain advantages and disadvantages for their use, they are fully explained. Included in this section is the

explanation for the reliance these measurement frameworks have on the reliability of the scores obtained. Since reliability estimates may differ between the two measurement theories, data characteristics affecting reliability estimates, such as sample size, number of items, type of response scale, and number of response choices are discussed.

The final section of this literature review focuses on the role sampling design, specifically multilevel modeling, plays in accurately estimating reliability coefficients across measurement frameworks (Feldt, 1990). A thorough discussion of the precision of reliability estimates in a multilevel model supports the need for and purpose of this current study. Lastly, reliability estimation procedures based on Cronbach's α (Cronbach, 1951) and polychoric ordinal α (Bonanomi, Nai Ruscone, & Osmetti, 2012; Zumbo, et al., 2007) used within the CTT and MCFA frameworks, as well as the person reliability used in the Rasch rating scale model (RSM; Andrich, 1978; Masters, 1982) are examined in both single and multilevel models.

Fundamental Measurement

Recognizing the need to develop accurate and accepted measures of mental and social phenomena, Thorndike (1904) introduced students of the social sciences to what he called "mental measurement" (p. 3), which he adopted and modified from the physical sciences. He explained that in the mental sciences, as in the physical sciences, the need to measure "differences, changes, relationships or dependencies" (p. 5) is just as important but present what he termed "special problems" when human factors are involved because often judgments about what is being measured conflict. Thorndike posited that the scientific method of measurement in the physical sciences is based on fundamental mathematical measurement principles which were established to provide accurate and

consistent measurement of the object or attributes being measured. He developed a mental measurement scale which incorporated several of the fundamental measurement principles. These principles are conservation, transitivity, and unit iteration (Annenberg, 2012). In the physical sciences, conservation is the principle that an object or attribute maintains the same size and shape regardless of orientation. For example, a person's height remains constant whether her or she is standing or lying down. Transitivity means that when you cannot compare two objects or attributes directly you must compare them via a third object or attribute. In other words, if A = B and B = C, then A = C. Unit iteration refers to the determination of the correct unit of measurement which requires a deeper understanding of the attribute being measured. For example, with distance, height, or length, a linear measurement is appropriate and when measuring area, two-dimensional units are appropriate. Fundamental measurement in the physical sciences therefore requires a stability of measurement which can be expressed in comparable units of measurement. If these three principles of fundamental measurement hold true, the concatenation of like units is possible which are applied every day in the physical sciences to measure quantities such as weight, height, length, width, and depth (Lindquist, 1989).

Thorndike (1904) argued that the ability to quantify, and therefore, measure human behavior was simply a matter of interpreting the underlying mathematical concepts to non-mathematicians. He provided the example of measuring the spelling ability among 10-year old boys. In essence, if you were to develop a list of 50 or 100 spelling words, who is to say that spelling *certainly* is of equal difficulty to spelling *because*? This measurement therefore, requires judgment, which means that agreement

about ability must first be established. Thorndike then posited that measuring mental traits such as abilities, beliefs, or attitudes, required an underlying continuum in which to mark the appropriate observed level of the trait. Thorndike then went on to develop an objective scale of measurement for which judges could agree. His definition of objectivity included aspects of reliability and validity and laid the foundation for CTT.

Concurrently, Spearman, (1904) developed the framework for CTT where a theoretical true score and error were summed and linked to an observed score. Spearman's CTT framework relied heavily on the reliability of the scores in terms of the amount of error in the observed scores.

The ability to measure psychological and social phenomena in a meaningful way using fundamental measurement principles relied on the development of response scales and the establishment of levels of measurement to better classify, and therefore, identify stable variable characteristics, mentioned here only to illustrate the relationship between fundamental measurement principles and response scaling, and discussed in more detail in the next section. Thurstone (1928) demonstrated that attitudes could be measured in a similar manner as variables in the physical sciences by placing responses on a linear scale in order for researchers to make a "more or less type of judgment" (p. 529) on a given trait of interest. Likert (1932) introduced a simplified version of Thurstone's scaling method which addressed some of Thurstone's unverified assumptions in response scaling for measurement in the social and psychological sciences. These assumptions are discussed in the next section. Stevens (1946) described measurement as "the assignment of numerals to objects or events according to rules" (p. 677) and introduced four new scales or levels of measurement; quite controversial at the time, but still in use today:

nominal, ordinal, interval, and ratio. These levels of measurement provide a good starting point from which to choose the correct statistical methods for a given data set based on the level of measurement of both independent and dependent variables. Thurstone and Likert scales are ordinal scales which means the numbers represent a position or rank in a sequential response pattern. Guttman (1950) developed a cumulative scaling model where items are ranked from easiest to most difficult and agreement with any particular item implies agreement with the lower-difficulty items. Deviations from the ideal Guttman pattern are considered random errors (Guttman, 1950). This is extended to achievement tests with dichotomous (correct or incorrect) outcomes where the assumption is if the examinee can successfully answer items of X difficulty, s/he would be able to answer preceding items of lower difficulty.

Applying fundamental measurement principles of quantifying variables by placing their measurement on a continuum from least to greatest amount, in conjunction with use of various scaling procedures, which are described later in the chapter, allows parameter estimators to be computed with greater efficiency. These principles laid the theoretical foundation for later IRT models by expanding the definition of fundamental measurement to include, (a) measurement which is not derived from other measurements and (b) measurement which is produced by an additive (or equivalent) measurement operation (Luce & Tukey, 1964; Rasch, 1960). These definitions are discussed in more detail in the section on IRT.

Following is a discussion regarding the development of three progressive response scales and the establishment of levels of measurement to classify and identify stable variable characteristics in order to apply the fundamental measurement principles

of conservation, transitivity, and unit iteration to more precisely measure social and psychological phenomena.

Response Scaling

Response scaling is at the core of psychometric theory. Psychometric theory enables comparisons between individual test scores or individual item scores by scaling differences among individuals based on a specific phenomenon or attribute of interest (Wright, Gaskell, & O'Muircheartaigh, 1997). Scaling models are developed for "three related but distinct purposes" (McIver & Carmines, 1981, p. 8): confirmatory, exploratory, or parallel analysis.

Confirmatory analysis is used to test hypotheses. For example, a psychometrician may test the hypothesis that there is a single dimension of hope underlying mental health recovery. The scaling model is then used as a point of comparison to evaluate how well the observed data fit the specified model. Exploratory analysis is used to describe the underlying structure of data. For example, it can be used to determine whether scores obtained on a survey developed to measure levels of hope confirm a unidimensional or multidimensional scale. The purpose of an exploratory scaling analysis is not to test a hypothesis of dimensionality but merely to discover latent traits related to a construct of interest, such as hope, depression, or self-efficacy. Finally, parallel analysis is used as a benchmark for related measures. For example, after developing a unidimensional measure of hope, a psychometrician will assess evidence of concurrent validity by correlating the scores on their scale with scores on a similar measure of hope.

McIver and Carmines (1981) explained that "scaling models may be used to scale persons, stimuli, or both persons and stimuli" (p. 9). Three scaling methods are elucidated

below, followed by a discussion of the evolution of response scaling to encompass item response theory (IRT).

Thurstone (1928), Likert (1932), and Guttman (1950) all developed unidimensional scales to measure attitudes. Each new scale developer identified the strengths and weaknesses of previous scales and worked to extend their usefulness in measuring psychological and educational phenomena. Roiser (1996) argued that "most attitude measurement concentrates on attitudinal differences and is thus psychometric, whereas Guttman scaling investigates attitudinal consensus [patterns of agreement] and is thus more suitable for the study of social representations" (p. 11). She explained that these response scales can be extended beyond attitude to include scientific understanding of psychological and educational phenomena.

Thurstone Response Scaling

Thurstone (1928) devised a method of measuring attitudes along a continuum by counting the number of opinions either rejected or accepted by the respondent. For example, drawing from current events, one respondent may be more in favor of same-sex marriage than another respondent. The Thurstone scaling procedures provide a "more or less type of judgement" (Thurstone, 1928, p. 536) where these opinions are located on a stated continuum based on attitudes conveyed. His research in and development of The Law of Comparative Judgment led to the development of three methods of response scaling: paired comparisons, consecutive intervals, and equivalent-appearing intervals (McIver & Carmines, 1981). Thurstone's methods scaled stimuli and then persons (Salkind, 2010). Thurstone provided the attitude of pacifism as an example and described the steps involved in his scaling method. He developed a qualitative continuous measure

of pacifism in which he (a) clearly defined pacifism; (b) used a set of opinions as anchors; (c) explained that pacifism could be represented by a single point on that continuum; and (d) used a series of graduated statements selected by judges for their representation of a single point on the continuum between extreme pacifism and extreme militarism. When participants either endorsed or rejected each statement, Thurstone was able to assess the strength and direction of their attitude toward pacifism. Further, by dividing the continuum into class intervals, he demonstrated the ability to count the frequency of the data points at each interval, thereby describing a group of individuals by means of a frequency distribution as illustrated by Figure 1 below:

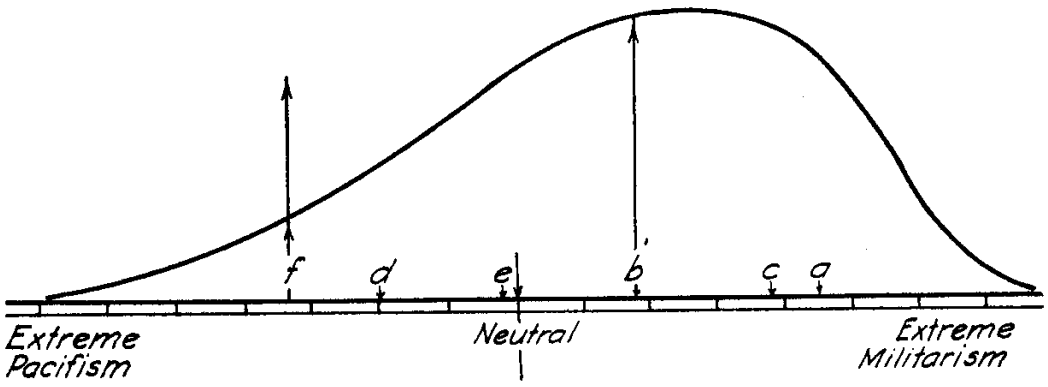

Figure 1. An example of Thurstone's (1928) continuum of summated group ability.

Thurstone's (1928) unidimensional method of scaling, while clearly defined, made several assumptions difficult to meet. Thurstone assumed that the aspect of measuring attitudes was "just as legitimate [as] to say that we are measuring tables or men" (p. 531). Next, he assumed the opinion of an individual was a statement of attitude and that individuals would be honest in their opinions. Finally, he had individuals write statements of opinion for a given variable of interest and used judges to specify the point

on the continuum represented by that statement. Two problems with this approach can be immediately identified: First, the judges were not selected based on specific expertise, but opinion; second, his method required hundreds of judges' opinions, making the approach impractical, which brings the scale into question. Regardless of these oft unmet assumptions, Thurstone made an important contribution to attitude response scaling and the ability to measure psychological phenomena.

Likert Response Scaling

Likert (1932) addressed some of the issues involved in social and behavioral measurement and developed a summative scale to address disadvantages in Thurstone's (1928) scaling. For example, he discussed at length the number of unverified assumptions included in Thurstone's attitude scales such as the independence of the scale values from the distribution of attitudes of the readers and the use of judges to correlate responses. Likert (1932) emphasized the need to simplify what he referred to as "exceedingly laborious" (p. 6) methods. Building on the social and psychological research of Thorndike (1904, 1913, 1918), Moore (1925), Allport (1929), G. Murphy (1929), and G. Murphy and Murphy (1931), Likert emphasized the role of theory in social and behavioral research and introduced the use of a five-point response scale. His scaling methods were used to scale subjects based on a single stimulus (level of agreement or disagreement). His unidimensional scale involved "a series of propositions to be responded to by the words *strongly approve, approve, undecided, disapprove, and strongly disapprove"* (p. 14). In order to quantify and measure the responses, Likert coded and ordered them from 1 to 5 with 1 = *strongly disapprove* to 5 = *strongly*

approve, with higher scores indicating more of the trait being measured, and summed the scores across all the items.

Likert advanced two main assumptions of his scaling model: the data of the summed scores are normally distributed and equal intervals of the ranges existed. To support the assumption of normality, Likert (1932) stated that "it seems justifiable for experimental purposes to assume that attitudes are distributed fairly normally and to use this assumption as the basis for combining the different statement" (p. 22). To support the assumption of equal intervals, Likert posited "[the scale] retains most of the advantages present in methods now used, such as yielding scores the units of which are equal throughout the entire range" (p. 42). Developed as an ordinal scale, these data are often treated as interval level data due to the summation of responses across all items which provides a total score. Additional research supports the underlying assumptions of normality and equal intervals of Likert's 5-point response scale which enables the use of parametric statistical tests such as t-tests and ANOVAs to analyze data (Allen & Seaman, 2007). Likert's 5-point response scale and its variations (Likert-type response scales: scales with fewer than or more than five categories), visual analog scales, and response scales based on anchor points rather than levels of approval) have been the primary scales used in survey research and self-report measures since their introduction in 1932.

Guttman Response Scaling

Guttman (1950, 1967) developed a scaling technique to be used as an alternative to Thurstone or Likert response scaling where a series of statements of attitude characterizes a progressively larger (or smaller) proportion of the population. For example, "a person who endorses the most demanding item should also endorse the most

consensual" (Roiser, 1996, p. 14). In other words, if the assumption of a Guttman (1950) scale is met (i.e., unidimensionality of the scale) and an individual endorses the most difficult item, then all items prior will also be endorsed. Therefore, the Guttman scale can be characterized as a cumulative scale, suggesting that the variation in the proportions of agreement, avoided in Thurstone or Likert scales, is at the heart of measurement, where the actual number of items endorsed is the recorded score as illustrated in Table 1.

Table 1

An Example of a Consistent Guttman Cumulative Scale

Person	Item 1	Item 2	Item 3	Item 4	Item 5
1	Yes	Yes	Yes	Yes	Yes
2	Yes	Yes	Yes	Yes	No
3	Yes	Yes	Yes	No	No
4	Yes	Yes	No	No	No
5	Yes	No	No	No	No

Note. Yes signifies the endorsement of the item. Adapted from Oppenheim, 1986, p. 147.

Roiser (1996) pointed out two critical differences between the Guttman scale and the Likert (or Thurstone) scale:

> Similar Likert scores may be achieved by endorsing different selections of items, [whereas] individuals with the same score may not actually have the same attitudes and two individuals scoring equally on a Guttman scale must be in complete agreement both on the items that they endorse and reject. (p. 15)

Item-level data collected using Likert and Likert-type scales are polytomous in nature while item-level data collected using a Guttman scale are dichotomously scored. Although not perfect representations of psychological phenomena, Likert and Guttman

scales are still in use today and provide important information for survey researchers, clinicians, and educators. As previously evidenced, the type of scaling method used is a key ingredient in survey research, self-report measures, and the assessment of aptitude, attitude, and objective measures of phenomena. Another key component of measurement which is closely aligned with summative response scales is the level of measurement (nominal, ordinal, interval, and ratio) introduced by Stevens (1946, 1951). These levels of measurement are commensurate with the scaling method chosen.

Levels of Measurement

Stevens (1946) proposed definitive classes or levels of measurement based on the mathematical properties of the scales. He argued that four levels of measurement existed: nominal, ordinal, interval, and ratio "based on the empirical operations needed to create each type of scale" (p. 678). He defined each level both by its basic empirical operations and mathematical group structure and went on to discuss the type of statistical analysis appropriate at each level. Table 2 details Stevens' four levels of measurement, which are routinely used in modern psychometrics.

The development and maturation of response scales and the establishment of levels of measurement to accurately capture attitudes and attributes invites spirited debate in the field of psychometrics, but nonetheless is fundamental to the development and evolution of measurement frameworks such as CTT, MCFA, and the one parameter IRT model. These three measurement frameworks are discussed in detail later in this chapter.

Table 2

Stevens' Four Levels of Measurement

Scale	Basic Empirical Observations	Mathematical Group Structure	Permissible Statistics
Nominal	Determination of equality	Permutation group	Number of cases Mode Contingency correlation
Ordinal	Determination of greater or less	Isotonic group: Any monotonic increasing function	Median percentiles
Interval	Determination of equality of intervals or differences	General linear group	Mean Standard Deviation Rank order correlation Product moment correlation
Ratio	Determination of equality of ratios	Similarity group	Coefficient of variation

Note. Adapted from Stevens (1946).

Response Scaling

Scaling methods are not limited to survey research. Clinicians, educators, and applied researchers embrace various response scales to measure and assess aptitude, attitude, and objective measures of a variety of phenomena. It is important to note that Guttman and Likert scales are found most often in assessments of aptitude or attitude and are discussed in that realm here.

Dichotomous vs. Polytomous Response Scaling

Aptitude tests are more likely to have dichotomously scored items, for which an item can be marked as correct or incorrect. Attitude measures frequently follow a response pattern with items measuring beliefs and feelings, where the respondent may choose between dichotomously scored options such as true or false, yes or no, or agree or

disagree. More commonly used in the measurement of attitudes is a rating scale response where the respondent chooses from a range of ordered responses (polytomously scored options). Most contemporary researchers prefer providing more than two choices to measure the latent trait of interest and previous studies provide strong evidence that a respondent's ability to choose among a range of responses will provide more measurement information than just two choices (Bejar, 1977; Kamakura & Balasubramanian, 1989; Masters, 1988). Samejima (1977, 1979) advised that polytomous data increase the statistical information of a given item when compared to dichotomous data, and in fact, when polytomous data are artificially dichotomized, substantial information is lost. Since dichotomous and ordinal data are categorical in nature, the categories represent imprecise locations along a trait continuum. Polytomous response options provide an advantage since there are more response categories from which to choose, providing more information over a wider range of the trait continuum than the range offered by dichotomous response options (Ostini & Nering, 2006). The current study focuses on items with five ordered response choices developed by Likert (1932) since these are the most commonly used response rating scales in aptitude and attitude assessment.

The polytomously ordered categories discussed in this dissertation are characterized by thresholds, or boundaries, along an observed response continuum used to measure the latent trait of interest. These boundaries separate the various categories and as logic dictates, they always comprise one less boundary than category. For example, with three categories (Like, Neutral, and Dislike), there will always be one category defined by two boundaries and with five categories, there will be four

boundaries separating them (Ostini & Nering, 2006). With polytomous data, since the probability of a specific response in a given category reflects a respondent's observed level of a measured trait, psychometricians focus on these boundaries because the probability of responding within a category is governed by the characteristics of the two neighboring boundaries. For example, consider a unidimensional assessment of hope using a 5-point Likert-type scale (1 through 5), where 1 = *low levels of hope* and 5 = *strong levels of hope*. A Wright map is created to visualize the person and item data on the same metric. The left side of the Wright map locates the person ability measures along the variable *hope*, where persons are signified by the # symbol. The right side of the Wright map locates the item difficulty measures along the variable *hope,* where items are identified by item number. Higher scores indicate an increased level of hope. Items with low difficulty would be endorsed only by individuals with a low level of hope and items with high difficulty would be endorsed only by persons with the greatest level of hope. Figure 2 below is a Wright map representing data collected to measure the latent variable *hope* using the 12-item Snyder Hope Scale (Snyder, 1994).

Note that high positive thresholds indicate the lowest point at which a person with a high level of hope would endorse an item of high difficulty (e.g., item 5) and low negative thresholds would indicate the lowest point at which a person with a low level of hope would endorse an item of low difficulty (e.g., item 11).

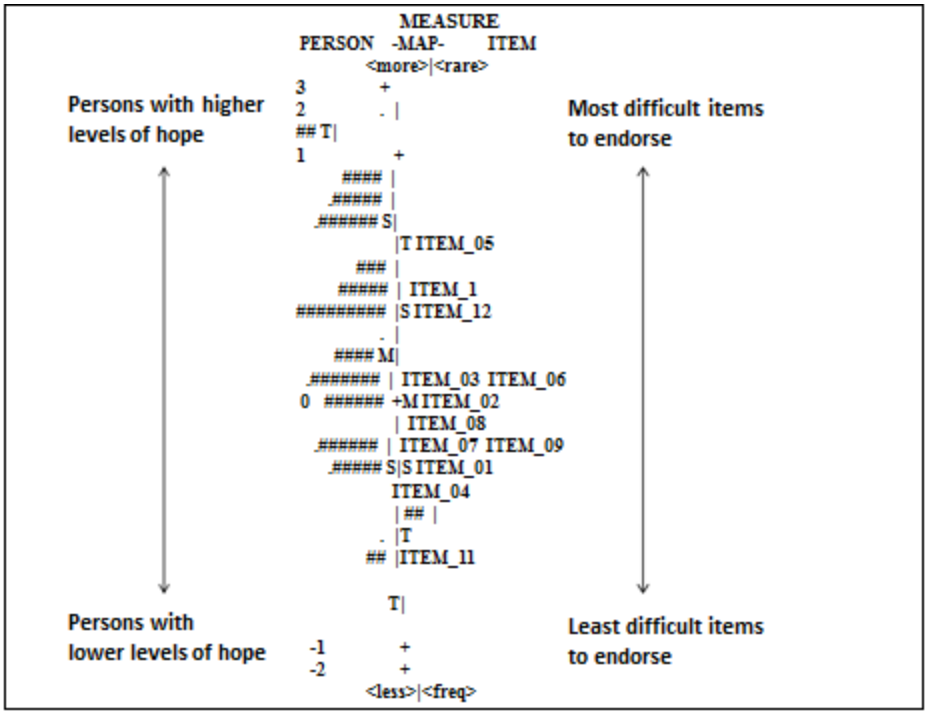

Figure 2. Wright map for a measure of Hope on a 12-item survey. Adapted from Linacre (2014)

Assessing individual and group responses, regardless of the response scaling method utilized or the way in which items are scored (dichotomously or polytomously), requires theories related to measurement. In the following section, I discuss the evolution of measurement theories and explore the advantages and disadvantages of each.

The Evolution of Measurement Theories

Introduction to Measurement Theories

In educational and psychological research, mathematical models are used to elucidate the underlying theoretical concepts of interest, provide a framework for comparisons, and define a context from which to conduct analysis and interpret results (Ostini & Nering, 2006). Mathematical models provide the means by which

psychometricians can quantify phenomena of interest. The simplest mathematical model is a count, for example, observing the number of times a given individual answers each test item correctly. For more complex analysis of a phenomenon of interest, more than one mathematical model is often employed to measure and assess underlying theoretical concepts of interest. Two such models are discussed at length here: Classical Test Theory (CTT) and item response theory (IRT). CTT is a set of mathematical models which evolved from research conducted by Spearman (1904) and builds on fundamental measurement described previously. IRT is an extension of CTT which allows for simultaneous measurement of person and item parameters.

Classical Test Theory

The early 20th century was a time of "exploration and [measurement] theory development" (Thorndike & Thorndike-Christ, 2010, p. 4) in the emerging field of psychology. Researchers began to recognize the existence of errors in measurement, understand errors as random variables, and conceptualize the idea of "correcting a correlation coefficient for attenuation due to measurement error [in order to] obtain the index of reliability" (Traub, 1997, p. 2). By differentiating between observed variable scores and error scores, the theory of measurement coalesced into what was known as true test theory, and finally regarded as CTT.

The framework of CTT was detailed by Spearman (1904) and others throughout the first half of the 20th century, culminating in the work of Lord and Novick (1968) and Allen and Yen (1979) regarding the use and analysis of mental tests and the precision of the test score (McDonald, 1999). Equation 1 is the true score model and is the basis of classical measurement theory:

$$X = T + E \quad (1)$$

The premise of CTT is that a given observed score *(X)* on a test is comprised of two components: a hypothetical true score (*T*) which represents the average score taken over recurrent independent testing, and random error (*E*). Therefore, the less random error in a given score, the better the raw score reflects the hypothetical true score. The true score of a person is found by taking the mean score that the person would get on the same test if he or she had an infinite number of testing sessions (or trials). The goal of CTT is to provide a framework to assess the observed score (*X*) of a test-taker by partitioning out the estimated random error (*E*) from the hypothetical true score (*T*). Allen and Yen (1979) explained that if the true and error score assumptions are met, and an individual were to take the same test 1,000 times, the average of the individual's raw scores would be the best estimate of the true scores. Furthermore, using the standard deviation of the distribution of random errors around the true score (known as the standard error of measurement) as an index, Allen and Yen demonstrated that if 1,000 people were to take the same test one time each, the true and error score assumptions are still met. This substitution simplified data collection and analysis enormously.

Assumptions of Classical Test Theory. Allen and Yen (1979) explained that the foundation of CTT was the idea of the true value of a variable (X_{True}). Classical Test Theory (CTT) assumes that the true values of scores on a variable, *X*, in a given population of interest follow a normal distribution denoted as *N*(0, 1). The observed distribution of the scores on the variable *X* is denoted as *D*. The population mean is denoted by μ and the population standard deviation is denoted by σ_{True}. Using this

notation, modified from Allen and Yen (1979), the distribution of the true value for a population of participants is found in Equation 2:

$$D(X_{True}) = N(\mu, \sigma_{True}) \quad (2)$$

The population parameters μ and σ_{True} differ from those used in a sample due to sampling error. CTT focuses on how the observed values of *X* (X_{obs}) are related to the true values of X (X_{True}). Since CTT purports that the observed values are a combination of the true values plus a component of random measurement error, CTT makes three assumptions about the error component:

1. The error component will have a mean of zero. Therefore, the observed mean will not be systematically distorted away from the true value by the error
2. The measurement errors are assumed to follow a normal distribution.
3. The measurement errors are uncorrelated with the true values.

Equation 3 represents the expression for the distribution of $X_{obs:}$

$$D(x_{obs}) = D(X_{True}) + N(0, \sigma_{err}) \quad (3)$$

where *D* is the observed distribution of the variable *X* and σ_{err} is the standard deviation of the normal random error term. Equation 4 shows that for an individual (*i*th) participant, the Equation 4 expression could be written as:

$$X_i = X_{i,True} + \varepsilon_{2i} \quad (4)$$

where $X_{i,True}$ denotes the value of X_{True} for participant *i*, drawn from the true value distribution $N_1(\mu, \sigma_{true})$, and ε_{2i} denotes the error term for the *i*th participant, drawn from

the error distribution $N_2(0, \sigma_{err})$. From these three assumptions, it follows that the expected value of the sample mean, (Exp{ $\bar{x}$ }) is μ . In addition, the sample standard deviation (s) of X_{obs} is going to be larger than σ_{True} as the random error component (with a standard deviation of σ_{err}) increases the variation in X_{obs}.

Equation 5 represents the expected sample variance (s^2) of a composite score. Imagine two variables *a* and *b*, and a variable *c (observed score)* which is the sum of *a* and *b*. The variance of the new variable *c* is given by:

$$Var\ (c) = Var\ (a) + Var\ (b) + 2r_{ab}*\sqrt{[Var(a) * Var(b)]} \quad (5)$$

where r_{ab} is the correlation between *a* and *b*

Since the observed values are the sum of the true values and random measurement error, using "true" and "error" instead of "a" and "b," the expected value of the sample variance is simply the sum of the variances of the true score and error terms. The final term in Equation 6 is absent because of the assumption that the measurement errors are uncorrelated with the true values. It is important to estimate the expected value of the variance in the observed score (s^2) in order to determine the amount of variance explained by the true score (σ_{true}^2) as seen in Equation 6:

$$\text{Exp}\{s^2\} = \sigma_{true}^2 + \sigma_{err}^2 \quad (6)$$

Advantages of Classical Test Theory. Classical Test Theory does not involve a complex theoretical model to assess (a) a test-taker's ability to correctly respond to a specific item (aptitude) or (b) to measure a specific attitude, but instead collectively assesses a pool of test-takers. As this dissertation is focused on measuring attitude,

ideally, the observed score reflects the test takers' true attitudes with minimal error. In other words, the observed score is similar to the theoretical true score. In this regard, ability refers to the ability to indicate more of the attitude being measured when the attitude is high. Lord (1953) explained that ability scores are test independent while observed and true scores are test dependent. In other words, test takers come to the test with a certain level of ability on the attitude being measured, while the observed and true scores will "depend upon the selection of assessment tasks [drawn] from the domain of assessment tasks over which their ability scores are defined" (cited in Hambleton & Jones, 1993, p. 253) In the case of measuring aptitude, CTT models the test-takers' proportion of correct responses to a specific item using dichotomous scoring. This is known as the P value of the item (not to be confused with the *p*-value as an indication of significance in hypotheses testing) and is used as the index for item difficulty, with lower values indicating a harder item and higher values indicating an easier item. P is the proportion of respondents who answer the item correctly. The ideal P value is .5, meaning that 50% of the test-takers endorse or pass the item, which J. B. Kline (2005) explains provides "the highest levels of differentiation between individuals in a group" (p. 96). More relevant to this dissertation is the case of measuring attitudes using polytomous rating scale-scoring. As with dichotomously scored items, polytomous items are used to quantify true score values on a trait of interest, defined here as the underlying ability of interest (the trait intended to be measured). As values of the true score increase, responses to items representing the same concept should also increase. In other words, there should be a monotonically increasing relationship between true scores and observed scores, assuming that responses are coded so that higher responses indicate more of the

measured trait. Item difficulty is represented as an index of the mean score of the item across test-takers (DeMars, 2010) with higher values indicating greater overall endorsement of the attitude trait.

Another important characteristic of items is discrimination. A higher discrimination index indicates that the item differentiates well between test-takers with different levels of the construct being measured. For example, in aptitude testing this means that the item discriminates well between test-takers of low and high ability. In attitude assessments, this means that the item discriminates well between test-takers with more versus less positive attitudes regarding the trait being measured. Therefore, high discrimination is preferred since it means the item, test, or measure is able to differentiate between those who know the material and those who do not or those with positive attitudes and those with negative attitudes. When an item discriminates well between higher and lower ability (or attitude) test takers, the relationship between the test taker's score and the overall scores on the test will increase. For polytomously scored items, the item discrimination value is computed using the Pearson product-moment correlation coefficient. For dichotomously scored items, point-biserial correlation is computed. When the correlation is positive, individuals who endorsed (or answered correctly) the item "score higher on the sum of the remaining items" (DeMars, 2010, p. 5) than those who do not endorse the item (answer incorrectly).

Disadvantages of Classical Test Theory. The main disadvantage of CTT is that item statistics are sample dependent and examinee characteristics (such as ability) are item-dependent. Fan (1998) described this as "circular dependency" (p. 1) in that not only are the true scores (person parameters such as ability) test dependent, the item difficulty

and item discrimination values (item parameters) are sample dependent (Lance & Vandenburg, 2010). For example, the unidimensional measure of the attitude *hope* contains items easy for a hopeful person to endorse (i.e., I set attainable goals), but difficult items for a less hopeful person to endorse. To summarize, the observed scores for the more hopeful person will increase and the observed scores for the less hopeful person will decrease. In other words, a hopeful person's ability estimate will increase with items considered more difficult for a less hopeful person to answer. Conversely, a less hopeful person's ability estimate will decrease because he or she is less able to endorse a more difficult item (i.e., I am never concerned about the future). Comparing true scores across tests would be difficult due to the differences in test properties. Additionally, item discrimination will be higher in samples that represent a large range of abilities. Finally, the item difficulty parameter depends upon the ability level of the sample. For example, if an exam regarding the Central Limit Theorem were given to fifth graders and the same exam to statistics majors in college, the item difficulty indices would vary substantially because what is hard for the fifth graders to conceptualize may be easy for the statistics majors.

Item parameter estimation (i.e., item difficulty and discrimination) is certainly an important disadvantage of CTT since these parameters are test and sample dependent which limits the generalizability of the results. To overcome this disadvantage, Thurstone (1928) proposed absolute scaling, which is an empirical, ad hoc procedure to measure invariance, and more commonly referred to within the IRT framework. The method employs standardizing scores so that the same metric is used to assess a respondent's location within the distribution of test scores.

Other disadvantages of CTT include the assumption that the standard error of measurement (SEM) of scores from a given test is equal across an entire population (Spearman, 1904). The SEM is frequently used to interpret individual test scores but is only useful if the test scores demonstrate high reliability and the obtained score for the individual test taker does not deviate significantly from the mean test scores of all test-takers. J. B. Kline (2005) explained that this means the "standard error does not differ from person to person but is instead generated by large numbers of individuals taking the test" (p. 94). For example, regardless of the magnitude of the observed score, the standard error for each examinee is assumed to be the same, which is unrealistic (J. B. Kline, 2005). In CTT, the standard errors for all examinees are expected to cancel each other out and therefore, sum to 0 (Lord, 1953). However, it is important to note that test-takers with the same total score may have different standard errors and that raw score standard errors are larger for overall scores closer to the mean than for extreme scores (Brennan & Lee, 1999) Finally, when the assumptions of CTT are not met, researchers may "convert scores, combine scales, and do a variety of other things to the data to ensure an assumption is met" (J. B. Kline, 2005, p. 94). Kline described the manipulation of data as problematic because of the possibility of ignoring systematic error. However, CTT is based on three parameters, observed score, true score, and error and most analysis conducted within the CTT framework is based on summing the observed scores across items, reducing error, and estimating true scores based on the model.

Estimating reliabilty in Classical Test Theory. Reliability is the overall consistency of the observed scores of a measure and the three most commonly used estimates of reliability in CTT are:

1. Test-retest reliability refers to the consistency of scores when the same test is given to the same people at different times (Nunnally & Bernstein, 1994)

2. Parallel-forms reliability refers to the consistency of scores when different people receive more than one form of a test measuring the same construct (Nunnally & Bernstein 1994).

3. Internal-consistency reliability refers to the consistency of scores across items (Cronbach, 1951).

Lord and Novick's (1968) defined reliability within the CTT framework as the ratio of true score variance, $\sigma^2_{T,}$ to the observed score variance, $\sigma^2_{X,}$, where the reliability of the observed test scores, *X*, is denoted as ρ^2_{xt} (see Equation 7). Pickering (2001) demonstrated the conceptual model of reliability based on computing the proportion of true score variance relative to total variance in Equation 8:

$$\rho^2_{XT} = \frac{\sigma^2_T}{\sigma^2_X} = \frac{\sigma^2_T}{\sigma^2_T + \sigma^2_X} \quad (7)$$

$$\text{Reliability} = \sigma_{True}{}^2 / (\sigma_{True}{}^2 + \sigma_{err}{}^2) \quad (8)$$

Two Coefficients to Estimate Internal Consistency

While several coefficients to estimate reliability within the CTT framework have been developed, this dissertation focuses on two coefficients to estimate internal consistency for polytomous data: Cronbach's coefficient α (Cronbach, 1951) and polychoric ordinal α (Gadermann et al., 2012; Bonanomi et al, 2012; Zumbo et al., 2007).

Cronbach's coefficient alpha. In the CTT framework, Cronbach's coefficient alpha (α) is the most frequently reported reliability coefficient for summated scales using

polytomously scored items (i.e., Likert or Likert-type scales; Hogan, Benjamin, & Brezinski, 2000). Developed by Cronbach (1951) to address the issues of simple split-half reliability examined by Spearman (1910) and W. Brown (1910), Cronbach's α is a maximum likelihood (ML) estimator of the parameter. In other words, the reliability of the scores cannot be less than the value of this parameter (Zeller & Carmines, 1980). Van Zyl, Neudecker, and Nel (2000) explained that Cronbach's α is equal to reliability under the assumption of tau equivalence; otherwise it is used as a lower bound estimate of the reliability of scores obtained on an assessment. The assumption of tau equivalence is addressed under the section on assumptions of coefficient α. Cronbach's α is a function of the number of items on a given assessment, the average covariance between item-pairs, and the variance of the total score. It can be viewed as the average correlation of a set of items measuring a specific construct or dimension of a construct. The coefficient α is defined in Equation 9:

$$K/(K\text{-}1)\ [\ 1\text{-}\Sigma\sigma_k^2/\ \sigma_{total}^2] \qquad (9)$$

where K represents the number of items; $\Sigma\sigma_k^2$ represents the sum of the variance of scores on each item and σ_{total}^2 represents the total variance of overall scores. Furthermore, van Zyl et al. (2000) explained that the ratio of variances expressed by Cronbach's α follows the general linear model (GLM) and as shown in Equation 9, Cronbach's α is item dependent. In other words, if the number of items increases, Cronbach's α will increase, and conversely, with fewer items Cronbach's α will be lower when holding all other factors constant. In addition, if the number of items is held constant, and the average inter-item correlation is low, Cronbach's α will change as a function of sample size. As

the average inter-item correlation increases, Cronbach's α increases (Cronbach, 1951; van Zyll et al., 2000).

Cronbach (1951) demonstrated that the coefficient α is "the average of all possible split-half coefficients for a given test" (p. 310). Cronbach's α is more frequently used to assess the reliability of scores obtained from polytomously scored items such as Likert response scales. Recent research suggests the fallacy of relying on Cronbach's α when polytomous data are used (Pastore & Lombardi, 2014; Rodriguez & Maeda, 2006; Sijtsma, 2009; Tavakol & Dennick, 2011; Teo & Fan, 2013; Zumbo et al., 2007).

Assumptions of coefficient alpha. Cronbach's α is rooted in two important assumptions:

1. Cronbach's α assumes unidimensionality of the measure, where all items measure the same underlying construct or latent trait. If the assumption of unidimensionality is violated, Cronbach's α will underestimate the reliability of the scores obtained (Geldhof, et al., 2014; Pastore & Lombardi, 2014; Rayvok & Penev, 2010; Rodriguez & Maeda, 2006; Sijtsma, 2009; Tavakol & Dennick, 2011; Teo & Fan, 2013; Zumbo et al., 2007).

2. Cronbach's α is grounded in an *essentially tau equivalent* model. This means that each item measures the same latent variable on the same scale with the same degree of precision, but that the individual item error variances are allowed to differ from one another, suggesting it is possible for each item to have its own amount of random error. This translates to all variance unique to a specific item is assumed to be the result of error (Raykov, 1997a, 1997b).

Advantages of coefficient alpha. Cronbach's α has three main advantages. First, it is included in all contemporary computer-based statistical packages such as SPSS, SAS, and R and therefore, is available to researchers across a wide range of academic fields. Second, it is a single measure of inter-correlations between items on a continuous scale, only requires one test administration, and may be more easily conceptualized by researchers than other estimates. Third, it is the most frequently reported reliability estimate in the world. Consequently, literature citing Cronbach's α across a variety of academic fields is easy to find.

Disadvantages of coefficient alpha. Cronbach's α has several distinct disadvantages. The first disadvantage is related to the standard error of Cronbach's α (SE_α), which provides an estimate of the amount of error found with the given scores. In turn this shows the spread of the inter-item correlations (Duhachek, Coughlan, & Iacobucci, 2005). The SE_α is inversely related to sample size and as stated by Duhachek et al. (2005), "heterogeneity within the covariance matrix negatively impacts reliability" (p. 299). Therefore, as the SE_α increases, reliability decreases (Cortina, 1993; Hattie, 1985; Schmitt, 1996). In the simplest case where all inter-item correlations are equal to the average of inter-item correlations (r), Cronbach's α can be expressed as Equation 10:

$$\alpha = \frac{kr}{1+r\,(k-1)} \tag{10}$$

and the standard error of α is expressed as Equation 11:

$$(SE\alpha) = \sqrt{Q/n} \tag{11}$$

where *k* is the number of items, *n* is the sample size, and Q represents the maximum likelihood estimator of alpha based on a standard assumption of multivariate normality (van Zyl et al., 2000). As n to ∞, $\sqrt{n}(SE_\alpha - \alpha)$ is normally distributed with a mean = 0 and variance represented by Q (described above) in Equation 12:

$$Q = \frac{2k(1-r^2)}{(k-1)[1+r(k-1)]^2} \quad (12)$$

The importance of the Equation 12 is that a confidence interval for Cronbach's α can be derived using the SE_α, which provides more information than a simple point estimate regarding reliability. As described by Equations 10, 11, and 12, the importance of considering the SE_α cannot be ignored, yet is rarely examined, calculated, or reported by behavioral, educational, and social science researchers (Cortina, 1993; Hattie, 1985;, Schmitt, 1996). Second, as Duhachek et al. (2005) and van Zyl et al. (2000) suggested, from these equations, it was clear Cronbach's α was both dependent on the number of items in an assessment (*k*) and the sample size (*n*). This meant that *k*, *n*, and *r* have a noticeable (negative or positive) effect on Cronbach's α, and researchers can affect inter-item correlations which in turn affects α simply by changing (increasing or decreasing) *k* or *n*. The effects of *k* or *n* are discussed in detail later in this chapter.

Third, many researchers wrongly assume that Cronbach's α is a measure of unidimensionality of a scale and do not understand the relationship among Cronbach's α, inter-item correlations, and SE_α. For these reasons, it is more often misinterpreted and over-utilized by well-meaning researchers (Cortina, 1993; Schmitt, 1996; Sijtsma, 2009). This single coefficient then takes on inflated meaning when it comes to making decisions regarding assessment development and analysis of the scores. If Cronbach's α must be

used, assessing and reporting confidence intervals for Cronbach's α would help guide these decisions (Sijtsma, 2009).

Fourth, since Cronbach's α is the default coefficient in statistical packages commonly used by researchers and frequently reported (correctly or incorrectly) in scholarly journals, it has become the "go-to" coefficient for reliability estimation, even when other reliability estimators would be more suitable based on the type of data and level of measurement (Sijtsma, 2009). Relevant to this dissertation is that over the past 20 years, researchers have provided compelling evidence that Cronbach's α is not appropriate for polytomous data (Bentler, 2009; Duhachek et al., 2005; Kopalle & Lehmann, 1997; Schmitt, 1996; Liu, Wu, & Zumbo, 2009; Sideridis, 1999; Sijtsma, 2009; Yang & Green, 2011; Zumbo et al., 2007), specifically Likert or Likert-type data such as those collected on a multi-item measurement such as a survey or attitude scale. Goodman and Kruskal (1979) and Norman (2010) disagreed that Cronbach's α was not appropriate for polytomous data and argued that even though the item responses are on an ordinal scale, the summated scores are on a continuous scale, which they felt suggested that Cronbach's α was an appropriate measure of internal consistency with polytomous data. There is, however, convincing evidence to the contrary. (Duhachek et al., 2005; Gadermann et al., 2012; Kopalle & Lehmann, 1997; Liu et al, 2009; Schmitt, 1996; Sideridis, 1999; Yang & Green, 2011). Since the calculation of Cronbach's α involves inter-item correlations, the Pearson covariance matrix is employed. In other words, "as measurement error increases, the observed inter-item-correlations will become more attenuated" (Fisher, 2014, p. 1). For example, a mental health client's score on an instrument measuring hope (where higher scores indicate more of the trait) may decrease

between the first and second administration of the instrument because she was just fired from her job. In this instance, the client's decreased hope score likely reflects measurement error rather than an underlying decrease in the trait of hope.

Consequences of Underestimating Alpha

Most relevant to this dissertation is that attenuated correlations will produce underestimated internal consistency reliability coefficients. Spearman (1904) explained that if reliability estimates are underestimated then those estimates would affect the correction for attenuation, which includes Cronbach's α since "measurement error refers to the inconsistency of measurement" (Fisher, 2014, p. 1). An important assumption for the use of a Pearson covariance matrix is that data are continuous. Violations of this assumption can "substantively distort . . . the [Pearson covariance matrix]" (Gadermann et al., 2012, p. 2). When data are from an ordinal scale rather than a continuous scale, the "desired distributional properties of continuous data" (Olsson, 1979, p. 443) are not present. Therefore, the evidence suggests that the Pearson correlation coefficient underestimates the true relationship between ordinal responses and the item inter-correlations (Haldago-Tello, Chacón-Moscoso, Barbero-García, & Vila-Abad, 2008). Cronbach (1951) discussed the difficulties in underestimating the coefficient when data lack variance.

When data are continuous, the numbers imply a proportionate rank order along a continuum, whereas, when data are polytomous, the numbers represent an ordered categorical label but do not necessarily have proportionate rank order (Rothke, 2010). Since variance is the average of the squared deviations from the mean, due to a restricted range of response choices, polytomous data cannot provide as much variance and,

therefore, may underestimate Cronbach's α (Nunnally, 1978). In addition, since the distance between 1 and 2 on a Likert or Likert-type scale may not be the same as the distance between 3 and 4 on that same scale, precise and meaningful measurement becomes more complex. For example, on an attitude scale where the item response choices are from 1 to 4, with 1 indicating a low level of the measured trait and 4 indicating a high level of the measured trait, two test-takers choosing the response option of 2 may differ in their actual level of the trait, with individual one considering the lower bound of 2 and individual 2 considering the upper bound of 2. Therefore, the number of response choices will substantially affect the variance of the scores obtained on each item.

Polychoric ordinal α. To address the misuse and underestimation of Cronbach's α when assumptions such as essential tau equivalence and/or unidimensionality are violated, Zumbo et al. (2007) tested a coefficient α for ordinal (polytomous) data. Known as the polychoric ordinal α, the coefficient uses the polychoric correlation matrix (Pearson, 1900; Zumbo et al., 2007), which takes into account the ordered categorical data structure rather than Pearson's correlation matrix, which assumes an interval level data structure (Haldago-Tello et al., 2008) and "severely underestimates the true relationship between two continuous variables when the two variables manifest themselves in a skewed distribution of observed responses" (Gadermann et al., 2012, p. 2).

The polychoric correlation matrix was proposed by Pearson (1900) where the measure of the relationship between two variables relies on the assumption of an underlying joint bivariate normal distribution and can be extended to ordinal data with a

joint normal distribution fundamental to his proposal (Pearson, 1907). In other words, the "polychoric correlation coefficient is the linear correlation of the postulated joint normal distribution" (Ekström, 2009, p. 3). The computation of this matrix is quite complex and beyond the scope of the proposed dissertation. The main differences between the Pearson correlation for continuous data and the polychoric correlation for ordinal data are the underlying distributions from which they are estimated. Both the Pearson correlation coefficient and polychoric correlation coefficient assume variables have an underlying bivariate normal distribution; however the polychoric distribution is based on the underlying latent continuous trait represented by the order categories while the Pearson correlation coefficient assumes a continuous standard normal distribution and represents the strength of the linear relationship between the row and column variables.

Advantages of polychoric ordinal α. There are three distinct advantages to using polychoric ordinal α for polytomous scales. First, conceptually, ordinal α is equivalent to Cronbach's α, but it is based on the polychoric correlation matrix rather than the Pearson correlation matrix. Therefore, empirical evidence suggests it is a more accurate estimate for measurements involving polytomous data (Gadermann et al., 2012; Zumbo et al., 2007). Second, polychoric ordinal α considers polytomous responses as expressions of the underlying latent trait and interprets the reliability of the observed ordinal scores using the observed item responses, where Cronbach's α interprets the reliability of the observed scores by treating them as continuous (Gadermann et al., 2012). Third, computer software packages such as SPSS (using the POLYMAT add-on), R, and SAS (using POLYCHOR) have advanced to the point that calculating or entering a polychoric correlation matrix to use in the polychoric ordinal α estimation can be accomplished and

the polychoric ordinal α coefficient is easily interpretable since the resulting metric is between 0 and 1 with 0 = no reliability and 1 = perfect reliability (Lewis, 2007; Zumbo et al., 2007).

Item Response Theory

Addressing the limitations of CTT to estimate item parameters that weren't sample dependent, and person parameters that weren't test dependent, item response theory (IRT) grew through the work of Richardson (1936), Lawley (1943), Lord (1952), Birnbaum (1957), Rasch (1960), Wright (1967), and Lord and Novick (1968). The focus of this dissertation regarding IRT models is the Rasch model, advanced in 1960 by George Rasch. Rasch developed a special case of the one-parameter logistic (1PL) IRT model to address the need for fundamental measurement principles in psychological measurement. Based largely on the work of Luce and Tukey (1964), the Rasch 1PL-IRT model places item difficulty and person ability on the same latent continuum by combining fundamental measurement with the composite theory of simultaneous conjoint measurement and continuous quantities to quantify psychological attitudes or attributes.

One of the assumptions for the Rasch IRT model is the responses across items should be uncorrelated, or locally independent, after controlling for person ability. For example, each endorsement or correct item response should be based solely on person ability and not on *trait* or *response dependence*, as explained by Marais and Andrich (2008). Marais and Andrich described local independence as being depicted in two ways. First, there may be *trait dependence*, where person parameters other than ability are part of the response (a violation of unidimensionality). Second, there may be *response dependence*, where the same person with the same level of ability has a response on one

item that depends on a response given for a previous item (a violation of local independence). For example, *trait dependence* is often found when tests are constructed to measure a single trait but the items are drawn from a test bank in which each item is intended to measure a different aspect of the trait of interest. *Response dependence* is found when a correct answer on a test provides a clue to the answers on successive items.

Luce and Tukey (1964) posited that simultaneous conjoint measurement is a new type of measurement that includes both fundamental and derived measurement. Fundamental measurement refers to measurement with "iterative unit values" (Bond & Fox, 2014, p. 15) such as weight and height, while derived measurement means that "the attribute itself (e.g., temperature and density) cannot be physically added together" (p. 16). Bond and Fox (2014) used weight, volume, and density to help readers conceptualize conjoint measurement in the non-physical world, such as measures of attitude and aptitude. In the case of weights, volume, and density described by Bond and Fox (2014), "the key to conjoint measurement does not reside in the collusion of fundamental measurement scales to produce a third derived measurement scale of density that conserves the crucial properties of scientific measurement already inherent in weight and volume" (p. 9). In other words, density is contained within weight and volume. According to Luce and Tukey, conjoint measurement can be seen as the observable relationships between and among the variable matrix cells.

Person and item characteristics are simultaneously (conjointly) measured and modeled by the Rasch model where person ability and item difficulty can be used to estimate the probability that a person of given ability will respond correctly to an item of a given difficulty (Rasch, 1960, 1977). Therefore, the independent variables of ability and

difficulty can be represented on an interval scale with common units of measurement as seen in Equation 13 (Bond & Fox, 2014):

$$P_i(\Theta) = \frac{e^{D(\Theta - b)}}{1 + e^{D(\Theta - b)}} \tag{13}$$

where θ is the person ability estimate, *b* is the item difficulty, and $P_i(\Theta)$ is the probability that a respondent of a given ability will respond correctly to an item of a given difficulty level (Rasch, 1960).

To illustrate Rasch's (1960) model of combining simultaneous conjoint measurement with concatenation, an example using dichotomously scored data collected on a subset drawn from the Geo-Science Concept Inventory (GCI v.1.0; Libarkin & Anderson, 2005) is provided. The GCI v.1.0 was developed to measure the latent trait of geo-science knowledge in topic areas such as earthquakes, volcanos, and plate tectonics. Each item on the GCI is scored as either correct (1) or incorrect (0).

Furthermore, a monotone transformation, or way of transforming the numbers representing correct and incorrect responses on the GCI v.1.0 into another set of numbers without losing the original order of the data, is accomplished in the Rasch model by using an inverse logistic transformation. For example, Equation 14 represents that for some monotonic transformation *M* (Perline, Wright, & Wainer, 1979):

$$M(P_{ij}) = \ln\left(\frac{P_{ij}}{1 - P_{ij}}\right) \tag{14}$$

where p_{ij} is the probability of a person (*i*) answering correctly to item (*j*) on the GCI v.1.0 and *ln* is the natural logarithm. That is, the Rasch model is additive in the person ability

(θ) and item difficulty (*b*) parameters which allows for practical applications in the estimate of these parameters (Perline et al., 1979).

The dichotomous Rasch model is presented here to provide background information and lay the groundwork for an extension of the Rasch model when scores are polytomously scored. One addition to the Rasch model is the rating scale model (RSM; Andrich, 1978) which Masters (1982), Wright (1984), and Andrich (1978, 2004) explained was an extension of the 1PL-Rasch IRT model to be used when data are polytomous and the same number of thresholds exist across items.

Rating Scale Model

The rating scale model (RSM) is a unidimensional model used to assess ratings with two or more ordered categories. RSM requires a fixed number of response categories for every item measuring the latent trait (Englehard, 2014). There are two different approaches to the RSM. Andersen (1977) introduced a response function, shown in Equation 15, in which the values of the category scores are directly used as a part of the function:

$$P_{ix}\,(\theta) = \frac{e^{w_i\theta - \alpha_{ih}}}{\sum_{x-1}^{m} e^{w_i\theta - \alpha_{ih}}} \tag{15}$$

where $w_1, w_2, \ldots, w_m$ are the category scores, or numeric values associated with each rating scale point, which prescribe how the *m* response categories are scored, and a_{ih} are item parameters such as item difficulty and invariance, connected with the items and categories. An important assumption of this model is that the category scores are

equidistant. In Andrich's RSM, item response functions to account for item thresholds are computed in Equation 16 as:

$$P_{ix}(\theta) = \frac{\exp \sum_{j=0}^{x} (\theta - (b_i + d_{ix}))}{\sum_{x=0}^{m} (\exp \sum_{j=0}^{x} (\theta - (b_i + d_{ij})))} \quad (16)$$

where d_{ix} is the relative difficulty of score category x of item i.

Assumptions of Rasch Item Response Theory and the Rating Scale Model

The Rasch IRT family of models has several strict assumptions: (a) unidimensionality of the test, (b) local independence, (c) nature of the item characteristic curve (ICC), and (d) parameter invariance.

1. Unidimensionality. As with CTT, unidimensionality requires that the items on a test or survey only measure one latent trait or construct.

2. Local independence. Local independence is the assumption that item responses are independent given a person's ability. Therefore, if person ability determines success on each item then ability is the only factor that systematically affects item performance. Once person ability is known (estimated), responses to items are uncorrelated.

3. Nature of the item characteristic curve. The logistic function specifies a monotonically increasing function so that higher ability results in a higher probability of success. In other words, item performance is regressed on the test-takers' ability. In addition, since the probability of endorsing an item is bounded at 0 and 1, the slope of the ICC captures the nonlinear relationship between item responses and the latent trait of interest.

4. Parameter invariance. Item and ability parameters do not vary over samples of examinees from the population of interest. In other words, two groups may differ in the distributions of the latent trait, but the same model should fit both.

Advantages of the Rasch Item Response Theory and the Rating Scale Model

The Rasch model of measurement is a special case of IRT. There are several advantages to using the Rasch family of models over CTT models. The Rasch model is based on estimating the probability of observing each response to an item as a function of ability on the latent trait being measured. Rasch modeling involves examining the probability of success (correct response) as a function of the item's difficulty and the person's ability. CTT is unable to separate person ability from item difficulty. Each item in Rasch IRT has its own item response function (IRF) represented by the item characteristic curve (ICC) which reflects item difficulty when ability is held constant. Therefore, an item's psychometric properties are taken into consideration by the model. Another advantage to the Rasch model is that it can be extended to polytomous data such as with the RSM. A third advantage, according to many researchers, is that Rasch is an excellent tool for evaluating construct validity and is invaluable in test development (Bond & Fox, 2014; Messick, 1989, 1996; Rasch, 1960).

Disadvantages of the Rasch Item Response Theory and the Rating Scale Model

The first disadvantage of fitting data to the Rasch model is the mathematical complexity of IRT models in general coupled with access to the software used in IRT. Applied researchers often lack training in measurement theories and rely on the more accessible tools developed for CTT. Another disadvantage is that the one-

parameter model (1PL) assumes that all items that fit the model have equivalent item discriminations. This is especially true with the Rasch model where all item discriminations are assumed to be = 1.0. Therefore, each item is only described by a single parameter (*item difficulty)* which the model assumes is the only item characteristic influencing performance. Finally, opponents of the Rasch family of models posit that these models are not robust to guessing, and instead consider guessing as a separate parameter. Proponents of the Rasch family of models explain that there are two types of "guessing." *random guessing*, which provides no information about item difficulty and person ability, and *informed guessing* which contains information about item difficulty and person ability. Smith (1993) provided several examples of how the Rasch model was able to detect *informed guessing* by assessing the person ability, item difficulty, the probability of answering an item correctly, and the response patterns of two individuals with a similar ability levels.

Estimating Reliability

The focus of this dissertation is to estimate reliability when data are polytomous in CTT, MCFA, and Rasch IRT frameworks. Reliability in CTT, MCFA, and Rasch theory "reports the reproducibility of the scores or measure, not their accuracy or quality" (Linacre, 2012, p. 23). In Rasch, two reliability estimates are calculated and each can range between 0 and 1, with values closer to 1 indicating higher reliability. The first is a *person reliability*, which is equivalent to score reliability in CTT. To achieve higher person reliability, a study must include either a person sample with a large range of ability and/or an instrument with many items. The second is *item reliability*, which is not reported in CTT but provides information about the consistency of the items and locating

the items on the latent variable (Boone, Staver, & Yale, 2014). To achieve higher item reliability, a study must include either an instrument with a large range of item difficulties and/or a large sample of persons. Reliability coefficients for all three measurement frameworks (CTT, MCFA, and IRT) are estimates of the ratio of *true measure variance* to *observed measure variance*. The height of the ICC can be used to assess item reliability. Linacre (2012) provided three rules of thumb for reliability estimates for Rasch models:

1. If the item reliability is less than .80, a bigger sample is required.
2. If the person reliability is less than .80, more items are needed in the test.
3. High item reliability does ***not*** compensate for low person reliability.

The Importance of Reporting Reliability Estimates

The idea of reliability in the context of educational and psychological assessments is mired in misunderstanding (Baugh, 2002; Coe, 2002; Nunnally, 1978, 1982; Thompson & Snyder, 1998). Often graduate students preparing themselves for a career in the educational, psychological, and social sciences, as well as some faculty members, scholars, educators, and researchers in these fields, erroneously consider reliability to be a stable attribute of a given assessment tool rather than dependent upon the scores obtained from the administration of these assessment tools (Thompson, 2003; Vacha-Haase, 1998). These scholars and leaders often fail to realize that reliability is not subsumed within the instrument but instead relies on the scores obtained using the instrument. This misunderstanding leads to anything from misinformation and the endorsement of meaningless assessments or interventions to improper high stakes decisions. A variety of methods have been developed to estimate the reliability of scores related to an assessment instrument within the CTT framework. These include inter-rater,

test-retest, parallel forms, split-half, KR-20 and internal consistency reliability coefficients (Clark, 2008; Cortina, 1993; Cronbach, 1951; Henson, 2001; Nunnally & Bernstein, 1994; Spearman, 1904). This dissertation focused on the measure of internal consistency credited to Cronbach (1951) since, of the aforementioned methods of reliability, Cronbach's α is said to be the most commonly used method of measuring reliability (Geldhof et al., 2014; Raudenbush, 1993; Raykov & Penev, 2010, and others). One reason is that Cronbach's α can be calculated from a single test administration, which saves both time and money over other methods requiring more than one administration of a test (i.e., test-retest and parallel forms). Cronbach foretold that his internal consistency coefficient (Cronbach's α) "is a tool that we expect to become increasingly prominent in the research literature" (Cronbach, 1951, p. 299). His prediction has certainly come true. However, since reliability is a characteristic of the scores obtained from an assessment tool rather than a number assigned to the assessment tool for all time, applied researchers in the educational, psychological, and social fields often do not understand the impact low reliability of the scores on a given assessment has on other results, which is discussed in more detail below.

In 1999, The American Psychological Association (APA) Task Force on Statistical Inference (Wilkinson, 1999) published recommendations for appropriately reporting statistical results in scholarly research. One of these recommendations emphasized the need to include estimates of reliability of the scores obtained from a given educational or psychological assessment. Underlying this recommendation was the understanding that "score unreliability attenuates detected study effects" (Hogan et al., 2000, p. 524). The APA taskforce explained the importance of remembering that a test is

neither reliable nor unreliable. Reliability is a property of the scores on a test for a particular population of examinees (Feldt & Brennan, 1989). Thus, authors should provide reliability coefficients of the scores for the data being analyzed even when the focus of their research is not psychometric

While anecdotally, researchers appear to rely upon Cronbach's α as a measure of internal consistency in an overwhelming number of articles, no empirical evidence regarding the frequency of use has been provided for more than 13 years. Hogan et al. (2000) and Charter (2003) addressed this issue of the frequency of use of Cronbach's α in educational, social, and psychological research. Hogan et al. (2000) reviewed the number of times Cronbach's α, along with other methods of estimating reliability, was reported between 1991 and 1995. Employing a systematic sampling technique of every third entry from 37 scholarly journals published between 1991 and 1995, Hogan and his colleagues examined tests found in The *Directory of Unpublished Experimental Mental Measures*, Volume 7 (Goldman & Mitchell, 2008), *Tests in Print V* (L. Murphy, Impara, & Plake, 1999), *Tests: A Comprehensive Reference for Assessments in Psychology, Education, and Business* (Maddox, 1997), and the Educational Testing Service (ETS) *Test Collection* (e.g., ETS, 1995). They selected 696 out of 2,078 educational, psychological, and sociological tests and found that Cronbach's α was reported in 533 out of the 696 tests (66.5%). The next most commonly reported reliability coefficient was test-retest reliability which accounted for 152 of the 696 tests (19.0%) selected.

Revisiting the work of Hogan et al. (2000), Charter (2003) reviewed the literature regarding the frequency of use of reliability estimates. He gathered data from 2,733 test critiques, 8 journal articles, and 47 test manuals published between 1927 and 2001 with

92.5% of the data from the years 1960 to 1990 and found 937 reliability coefficients he deemed “sufficient enough to be used . . .” (p. 294). Table 3 shows the reliability coefficients used along with their frequency of use. Note that Cronbach’s α (Alpha) was used more frequently than any other method except test-retest. Charter acknowledged the discrepancies in the use of these various methods and explained that Hogan et al. (2000) used unpublished tests in 37 journals between 1991 and 1995 while he used mainly published tests from 1960 to 1990. An additional reason, which was not explored, is that data gathered by Charter include dates prior to Cronbach’s (1951) publication on reliability as a measure of internal consistency.

Table 3

Frequency of Reliability Estimation Methods Between 1927 and 2001

Method	Frequency	Relative Frequency
Alpha	140	14.94%
Alternate Forms	40	4.27%
Inter-judge	84	8.96%
KR-20	62	6.62%
Other or unknown	46	4.91%
Split Half	126	13.45%
Test-Retest	439	46.85%

These gaps in the research literature pose serious issues regarding the use, and possibly misuse, of Cronbach’s α. For example, in the past 20 years, the use of Cronbach’s α has continued; however, (a) no comprehensive study has focused on either

the number of times Cronbach's α has been used in educational and psychological research, (b) the number of researchers overlooking reliability estimates when reporting results from their studies is unknown, and (c) information regarding data characteristics when using Cronbach's α to estimate reliability is not available.

Regardless of the methods of estimating reliability, failing to consider reliability evidence puts into question any interpretation of research results since reliability is not only affected by data characteristics but affects other statistical properties as well. These data characteristics and statistical properties are discussed in the next section.

Statistical and Psychometric Properties Affected by Reliability

Reliability is not only affected by data characteristics such as sample size, number of items on an assessment, number of response choices, and sampling design, which are discussed in detail in the next section, but affects other statistical properties such as effect size, validity, *p*-value, power, and Type II error. Each of these statistical and psychometric properties are discussed below as they are related to reliability. Whether as stand-alone statistics or when combined, these properties express meaningful results and allow for accurate inferences.

Effect Size

Effect size, also known as practical significance is independent of sample size and refers to the magnitude of the impact of one variable on another (Huberty, 2002). The two most common types of effect size are (a) the effect size which focuses on the standardized mean differences between groups (Cohen's *d*; Cohen, 1969, 1988) and (b) the effect size focusing on the amount of covariation between the independent and dependent variables (e.g., a squared multiple correlation, adjusted R^2, or η^2). Cohen's *d*

(Cohen, 1969, 1988), which is the standardized mean difference between groups, is shown in Equation 17:

$$d = \frac{M_{group1} - M_{group2}}{SD_{pooled}} \tag{17}$$

where the numerator is the difference between two group means and the denominator is the pooled standard deviation as described in Equation 18:

$$SD_{pooled} = \sqrt{(SD^2_{group1} + SD^2_{group2})/2} \tag{18}$$

where the standard deviations of both groups are summed and divided by two. If within-group variance is reduced, effect size increases (Zimmerman, Williams, & Zumbo, 1993). An example of the relationship between power and effect size is provided by Cohen (1988) in Equation 19:

$$\boldsymbol{ES = ESP}\left(\sqrt{\boldsymbol{rxx'}}\right) \tag{19}$$

where ES is observed effect size, ESP is the population effect size, and $r_{xx'}$ represents reliability. Therefore, when reliability is 1, the observed ES is equal to ESP; but when reliability is < 1, the observed ES is a value smaller than the true ESP. R^2 and η^2 measure the degree to which variability among observations can be attributed to the conditions or explanatory variables as represented in Equation 20 (Cohen, 1977; Huberty, 2002):

$$\eta^2 \frac{SS_{treatment}}{SS_{total}} \quad (20)$$

where $SS_{treatment}$ is the sum of squares for the treatment groups or other, non-treatment categorical variables and SS_{total} is the total sum of squares in the model. Thompson (1994) explained that "score reliability inherently attenuates effects sizes . . . [and] we may not accurately interpret the effect sizes in our studies if we do not consider the reliability of the scores" (p. 840). More recently, Baugh (2002), Coe (2000), Durlak (2009), Gerhart, Wright, McMahan, and Snell (2000), R. Kline (2009), Wilkinson (1999), and others have provided evidence that effect size reflects other characteristics of a study such as estimates of internal consistency reliability. Thompson and Snyder (1998) studied issues related to reliability in peer-reviewed educational and psychological research and found that;

> The concern for score reliability in substantive inquiry is not just some vague statistician's nit-picking. Score reliability directly (a) affects our ability to achieve statistical significance and (b) attenuates the effect sizes for the studies we conduct. In other words, because measurement error variance is generally considered random, measurement error inherently attenuates effect sizes. It certainly may be important to consider these dynamics as part of result interpretation, once the study has been conducted. (p. 438)

According to CTT, the "observed" score is comprised of a "true" score, together with a component of "error," which can be conceptualized as "augmenting and diminishing [observed values]" (Spearman, 1904b, p. 89). Therefore, the amount of variation in true scores in a given sample will depend on the variation of both observed and error scores. This fluctuation in variation affects both Cronbach's α and effect size. Poor reliability will yield low Cohen's *d* (Thompson & Snyder, 1998).

Validity

In the broadest sense, validity of scores obtained on an assessment tool refer to the degree to which scores measure the latent trait of interest. There can be no validity of scores without first achieving reliability. Scores can be consistent (reliable) but unless they reflect what is actually being measured, the scores may not be valid (Moss, 1994; Weiner, 2007).

P-value

The *p*-value is used in hypothesis testing and represents the probability of obtaining the observed effect (or larger) under a null hypothesis, or hypothesis of no effect or difference. Ideally a *p*-value refers to the degree to which the results obtained by the sample are representative of the population, unless the sample contains bias. Therefore, a small *p*-value (i.e., under a given threshold of .05 or .01) indicates that the observed effect is not likely to have happened by chance and provides statistical evidence against the null hypothesis. Therefore, a low Cronbach's α indicates more measurement error which translates to a higher *p*-value (J. B. Kline, 2005).

Power

Power of a statistical test (1-β) refers to the ability to detect group differences or relationships between variables when they actually exist. In other words, the power of a statistical test is the probability that the null hypothesis was correctly rejected. Power is expressed between 0 and 1, with numbers closer to 1 indicating more power. Therefore, as power increases, the probability of a type II (β) error decreases. Power analysis can be used to calculate the minimum sample size required in order to be reasonably likely to detect a given effect size and conversely, power analysis can be used to calculate a

minimum effect size one can expect from a given sample size. Reliability affects statistical power through effect size (refer to Equations 17, 18, 19 and 20 above). Since reliability is characterized by "observed variance in conjunction with true or error variance, power changes as reliability changes only if observed score variance changes simultaneously" (Zimmerman & Williams, 1986, p. 123). Additionally, "if true score variance remains constant but lower reliability leads to increased error variance, then statistical power will be reduced because of the increased observed score variance" (Kanyongo, Brook, Kyei-Blankson, & Gocmen, 2007, p. 83).

Type II Error

Type II error (β) refers to failing to reject the null hypothesis when in fact the null hypothesis is false. In other words, finding no difference or relationship when, in fact, there was a difference or relationship. Poor reliability could lead to decreased statistical power in the presence of increased observed score variance, which could lead to increased Type II error (Roxy, Olson & Devore, 2011).

The importance of accurately estimating and interpreting reliability coefficients whether within the CTT or IRT framework cannot be underestimated since all estimated reliability coefficients influence effect size, validity, *p*-value, power, and Type II error, and severely jeopardize results. Improperly estimated reliability coefficients will potentially introduce additional relative bias.

Factors Affecting Reliability

As mentioned previously, many factors affect the reliability of scores obtained from test-takers on a given assessment tool. These factors include data characteristics such as sample size, number of items, number of response choices, and sampling designs

(i.e., single-and multilevel sampling designs) often found in educational and psychological research.

This section is organized by first addressing factors affecting reliability regardless of the sampling design, then examining these factors through the lens of single-level modeling, and finally, through the lens of multilevel (e.g., two-level) models within CTT and Rasch IRT measurement frameworks. Single-level models focus on individual effects and multilevel models examine individual effects at level-1 and group effects at level-2 while allowing for residual components to be estimated at each level (Geldhof et al., 2014; Preacher, Wichman, MacCallum, & Briggs, 2012; Raudenbush & Bryk, 2002).

Single-Level Sampling Design

Sample size. It is well documented in psychometric analysis of single-level models that as sample sizes (n) increase, reliability estimates such as Cronbach's coefficient α, polychoric ordinal α, and person and item reliability increase (Raudenbush & Bryk, 2002; Zumbo et al., 2007). Large sample sizes provide more reproducible and stable estimates of reliability, regardless of reliability magnitude and are more easily interpreted (Charter, 1999, 2003; R. Kline, 2014; Linacre, 2014; and others). However, in extremely small sample sizes, standard errors are often underestimated, correlation coefficients are less stable and tend to be over-estimated, and outliers play a role, resulting in higher reliability coefficients (Frost, 2015). Therefore, reliability coefficients must be interpreted with caution, taking into account sample size, standard errors, and outliers. This is also true for the number of items and the number of response choices on an assessment tool. As the number of items and/or the number of response choices

increase, reliability increases. In other words, research shows that larger sample sizes produce larger estimates of reliability across a range of data characteristics.

For behavioral and educational research related to reliability estimation, sample size recommendations vary from 30 participants to over 1,000 participants (Charter, 1999, 2003; Draxler, 2010; S. B. Green, Akey, Fleming, Hershberger, & Marquis, 1977; Hattie, 1985; Kahn, 2014; Peterson, 1994; Rust & Golombock, 2008; Yurdugul, 2008; and others). There are currently no studies examining appropriate sample sizes for the polychoric alpha reliability coefficient. Zumbo et al. (2007) developed a study of the stability of polychoric ordinal α when compared to Cronbach's α, as described in detail previously, and recommended exploring the impact of sample size in terms of the precision of polychoric ordinal α estimates in future studies, which is one purpose of this dissertation. Since sample size affects Cronbach's α within the CTT framework as well as person reliability and item reliability within the Rasch IRT framework, the debate regarding both ideal and realistic sample sizes within these frameworks is examined (Charter, 1999, 2003; S. B. Green et al., 1977; Hattie, 1985; Linacre, 2014; and others).

Classical Test Theory. As discussed previously, among the most commonly reported reliability coefficient in CTT is Cronbach's α coefficient (α) which is one method of estimating internal consistency (reliability) explored in this dissertation and which is a maximum likelihood estimator of the parameter and follows the GLM. Nunnally and Bernstein (1994) only vaguely referred to reliability as a function of sample size by stating that "measurement theory is large sample theory" (p. 228) and indicated that in order to precisely estimate reliability, larger samples are required. They surmised that a minimum of 300 participants was necessary for accurate reliability analysis but did

not provide any evidence to defend their claim. Furthermore, when examining corrections for attenuation, Nunnally (1967) stated that 300 participants was "a relatively small number of cases" (p. 218). Since these statements appear contradictory, several studies were conducted between 1994 and 2014 to provide guidelines (Charter, 1999; Peterson, 1994; Segall, 1994) or empirical evidence using simulated data (Charter, 2003; Yurdugul, 2008) of necessary sample sizes to estimate reliability.

Guidelines for estimating reliability. Many studies provide guidelines for appropriate sample sizes to more accurately estimate reliability; however, only a handful of studies provide empirical evidence to support their recommendations. Peterson (1994) conducted a meta-analysis of the use of Cronbach's α across 832 journals and 4,286 articles between 1960 and 1992. The journals represented behavioral, educational, marketing, and social science research and the following sample sizes are reported in Table 4.

Table 4

Reported Sample Sizes to Estimate Cronbach's Alpha

Sample Size (*n*)	Number of Articles
< 100	1,028
100-199	1,169
200-299	696
> 300	1,265
Not reported	128

The average sample size in Peterson's analysis was 268 which is on par with Nunnally and Bernstein's (1994) recommendation of > 300; however, Peterson (1994) concluded that sample sizes > 100, which he based on actual sample sizes utilized rather than any empirical evidence, were appropriate. Segall (1994) proposed sample sizes > 300 but failed to provide an explanation. Each of these recommendations aligned with Nunnally and Bernstein's proposed sample size of > 300 but none presented adequate reasoning. The sample size debate continues and the traditional guidelines on the necessary sample size for accurate reliability estimation yielded to confidence interval and parameter estimation using real and simulated data. Charter (1999) conducted a study using a 95% confidence interval for test-retest, parallel forms, split-half, and Cronbach's α reliability coefficients provided in previous studies (Charter, 1997; Feldt, Woodruff, & Salih, 1987) to determine more appropriate sample sizes when estimating reliability. First Charter (1999) indicated the coefficient *r* (e.g., correlations of .5, .6, .7, .8, .9, and .95) at varying sample sizes *n* (e.g., 50, 100, 200, 300, 500, and 1,000). Second, he calculated the width of the confidence intervals, although he did not expand on the formulas used for this step. Third, the results were plotted on a graph where the X-axis represented the sample size *n*, and the Y-axis represented the width of the confidence interval. Charter (1999) found that when sample sizes were < 50, reliability coefficients had larger standard errors than when sample sizes were 300 (.0605 compared to .023, respectively) and smaller sample sizes severely underestimated Cronbach's α. He concluded that Nunnally (1967) and Nunnally and Bernstein's (1994) sample size recommendation of > 300 was "probably too low [and] these figures suggest that at high *r's*, say .9 or above, one should have a minimum of 400 subjects and strive for more" (p. 563). Charter (2003)

conducted further investigation into 937 reported reliability coefficients and compared them to reliability standards outlined by Cronbach (1951), Rosenthal and Rosnow (1991), Cicchetti (1994), and Nunnally and Bernstein (1994). Charter determined that while sample sizes < 100 were common, sample sizes > 500 provided higher reliability estimates which he deemed to be more accurate.

Yurdugul (2008) conducted a simulation study in which he examined eigenvalues to determine the appropriate sample size for estimating reliability. First Yurdugul generated population data by generating a multivariate normal distribution based on the Likert 5-point scale and using the bootstrapping method of sampling. Bootstrapping refers to a resampling method where random samples of the parameter of interest are drawn and replaced to provide a more precise estimate of the population parameter of interest. In Yurdugul's (2008) study, N = 10,000 which included N = 5,000 observations and varying numbers of randomly determined variables (e.g., number of items). Using principal component analysis (PCA: a method of data reduction) coefficient α and λ_i resulting from the PCA was calculated from each population data set but not reported as part of the analysis. The estimated α, λ_i, and eigenvalues (the magnitude of variance in the data) were examined. Yurdugul (2008) explained that since standardized alpha is based on a correlation matrix of item scores it is therefore, "directly related to the eigenvalue of the first un-rotated principal component" (p. 398). Second, for each population data set, 100 samples were drawn by simple random sample methods with replacement for sample sizes n = 30, 100, 300, and 500 and item numbers k = 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20. For each sample data set, bootstrap estimators of Cronbach's coefficient α were generated. Finally, the relative bias (R- bias) and the

relative root mean square error (R-RMSE) were computed for every 100 samples in each of the four sample size levels. R-Relative bias refers to the degree of non-response error and R-RMSE is a measure of error in the responses. PCA was then conducted on every combination of the data characteristics described above (e.g., sample size, number of items, estimated factor loadings) and Yurdugul found that the minimum sample size for Cronbach's α is dependent on the level of the largest eigenvalue obtained. Using data simulated as unidimensional and normally distributed, Yurdugul concluded that a sample size of at least 100 was sufficient to produce an acceptable unrelative biased estimator for Cronbach; s α.

The evolution of sample size determination based on the confidence interval, correlations, and eigenvalues provides more robust and defensible methods on which to make recommendations regarding adequate sample sizes to applied researchers than previous speculation. The idea that reliability is a function of sample size is well founded. While other researchers laid the foundation for determining sample sizes for more precise measurement, B. Muthén and Muthén (2000) contributed to the sample size debate within the framework of structural equation modeling (SEM) by tying sample size not only to the number of parameters being estimated, but considered the reliance of other data characteristics such as unidimensionality, distribution, missingness, the number of items, and the number of response choices.

The consensus on sample size recommendations for accurate reliability estimation within the CTT framework is somewhere between 100 and 500 and is dependent on the number of parameters to be estimated (e.g., number of items and number of response choices).

Rasch item response theory. Rasch IRT assesses reliability with an analysis of person reliability, person separation, item reliability, item separation, inter-item correlations, and item difficulty thresholds. Person reliability is the degree to which items distinguish students' ability levels in a consistent manner which is analogous to Cronbach's α. The number of levels of student ability found in the data is known as person separation. Item reliability depends on the item difficulty variance independent of test length. The number of levels of item difficulty found in the data is known as item separation. Together, person reliability, person separation, item reliability, and item separation form the basis of overall reliability of the scores. Scores are considered to have good reliability if person reliability, person separation, item reliability, and item separation values are high. Respondents' (or test takers) scores are considered to have poor reliability if the person reliability, person separation, item reliability, and item separation values are low (Bond & Fox, 2014). Inter-item correlation (ICC), not to be confused with item characteristic curve, represents the average correlation each item has with other items on an instrument. The Rasch model assumes unidimensionality which means that the inter-item correlations should be at least moderate to high. Item difficulty (also called threshold) is a value that indicates how easy or difficult an item is. Ideally, an instrument will include items that fall across a spectrum of difficulty levels (Bond & Fox, 2014; Linacre, 2014).

Guidelines for estimating reliability in Rasch. Recommendations for appropriate sample sizes in Rasch IRT are based largely on three important considerations: (a) the number of parameters being estimated, (b) whether the items reflect a single administration of a test or the items are calibrated as part of a test bank,

and (c) the data characteristics and IRT model chosen (e.g., one parameter, rating scale. or partial credit models). There is no study specifically focused on the impact of sample size on reliability estimates within the Rasch IRT framework, which is one focus of this dissertation; however, Wright and Stone (1979) explained that a minimum sample size for an exploratory study in a simple Rasch IRT is 30, and as with CTT, larger sample sizes produce more accurate parameter estimations. K. E. Green and Frantom (2002) recommended a Rasch IRT study have a sample size > 100 participants and Van der Leeden, Busing, and Meijer (1997) recommended that a study have a sample size > 30 participants. Linacre (1994) stated that parameter estimates in Rasch IRT analysis (including reliability estimates) will be less precise with smaller sample sizes and recommended a sample size of at least 50 for the most basic Rasch IRT analysis. Reeve and Fayers (2005) suggested that in the case of items being calibrated for a test bank, a sample size > 250 is required while Emberson and Reise (2000) recommended a minimum of 500 participants. When estimating parameters using the rating scale or partial credit models, Reeve and Fayers and Emberson and Reise agreed that a minimum of 250 participants are necessary for accurate parameter estimation while Thissen and Wainer (2001) believed a sample size of 500 to be too low. None of these researchers provided any evidence supporting these recommendations beyond previous researchers' claims and trial and error. In this dissertation I examined the impact of sample size on reliability estimation in a rating scale model for both single and multilevel models.

The consensus on sample size recommendations for accurate parameter estimation in Rasch IRT is somewhere between 30 and 500 depending on the use of the items (e.g., single administration or test bank), the number of items, and the number of

response choices (e.g., dichotomously scored items versus items using a rating scale or partial credit model). As with CTT, more items on an instrument require larger sample sizes. One noticeable difference between CTT and Rasch IRT sample size discussions is that parameter estimation is thought to be more robust in Rasch IRT than CTT in the case of smaller sample sizes (K. E. Green & Frantom; 2002; Linacre, 2012; Wright & Stone, 1979).

Number of Items

In both CTT and Rasch IRT frameworks, the number of items affects the reliability coefficients calculated from the data (Cortina, 1993; Crocker & Algina, 1986; Jackson, 2003; Linacre, 2014; Nunnally & Bernstein, 1994 and others). As evidenced by Equation 9 shown previously, Cronbach's α estimation relies on the number of items and the mean inter-item correlation. Therefore, as the number of items increases, reliability will increase, but this concept can be misleading since a large number of items (e.g., items, 50) will produce higher reliability estimates even if the underlying inter-item correlations are small to moderate (Nunnally, 1978). B. Muthén, (1981) explained that the more items selected to measure a latent trait of interest, the more complex the model. Hellman, Fuqua, and Worely (2006) conducted a reliability generalization study and determined that above and beyond sample size, the number of items used to measure a latent trait was significantly correlated with reliability with more items resulting in higher estimates of reliability. Churchill and Peter (1984) conducted a meta-analysis regarding data characteristics such as sample size, response rate, sample demographics, and number of items and the effect each characteristic had on reliability estimates. One-hundred-fifty studies were included in the assessment on the effect of the number of items on reliability

estimates (Cronbach's α, test-retest, parallel reliability). The average number of items reported was 13.5 with a range of 1 item to 103 items. Churchill and Peter concluded that higher reliability estimates were significantly correlated with higher number of items, which is supported by the reliability generalizability study conducted by Hellman et al. (2006). While the literature is abundant with study results providing basic guidelines for the appropriate number of items to include in a measurement instrument (Allen & Yen, 1979; Draxler, 2010; Nunnally, 1978 and others), only a handful of empirical studies regarding the number of items and their specific effect on reliability estimates exist. Nunnally (1978) pointed out that the fallacy is that more items will increase reliability when in actuality, the combination of higher numbers of items and higher covariance among the items, the more reliable the scores In other words, since correlation is a scaled version of covariance, adding items (for the sake of having more items) that have little to do with the latent trait of interest will decrease covariance as well as Cronbach's α because when the sum of the variance is 0 Cronbach's α is also 0. In determining the number of items appropriate to include in a measurement instrument, I examined dozens of recommendations across a variety of disciplines and found that most recommendations are based on what "seems appropriate" and not on empirical evidence. For example, Draxler (2010) recommended between 9 and 45 items are required to measure one latent trait of interest while Karabatsos (2000) advised 20 to 50 items. Allen and Yen (1979) recommended writing "one and a half to three times" (p. 118) the number of items required to adequately measure the latent trait of interest and then conduct a pilot study using all of the items with a minimum of 50 participants and "item analysis procedures [to] identify poor items" (p. 118). One empirical study conducted by Jenkins and Taber

(1977) used simulated data to specifically address and assess the effects of various data characteristics, to include number of items, on reliability estimates. Building on the work of Lissitz and Green (1975) regarding the number of response choices and the effect on reliability estimates, Jenkins and Taber varied the number of items as well as other data characteristics in their Monte Carlo simulation study to assess the effects on reliability estimates. Data were generated to reflect a composite scale using a multivariate random number generator with seven levels of items numbers (2, 3, 5, 7, 9, 10, 14), seven levels of response choices (2, 3, 5, 7, 9, 10, 14), and three levels of covariance among items (.2, .5, .8). This resulted in a 7 X 7 X 3 fully crossed design. Their results suggested that as the number of items increased in the composite scale, the reliability estimates increased, especially when the error variance in individual item scores were high. While no specific recommendations regarding the number of items were provided, their findings regarding the number of response choices necessary to produce a stable estimate of reliability supported Lissitz and Green's (1975) study and are discussed in the section under response choices.

Finding little guidance in the reliability literature and to determine an appropriate number of items needed to measure a latent trait, previous studies following the GLM and using ML estimators commonly used in CFA were examined. The reasoning for this decision is three-fold: first, a scale's internal structure has implications for reliability as well as validity since it reflects "internal consistency by revealing which items are consistent with which other items" (Furr & Bacharach, 2014, p. 331) and CFA and SEM are used to assess internal structures of the data collected. Second, although not commonly used to assess reliability, factor analysis, including PCA, is considered by

some researchers to be an adequate alternative which may provide greater understanding of reliability (consistency; Raykov, 1997b; Yurdugul, 2008). Third, a CFA-based reliability estimation procedure for a unidimensional scale has been developed and provides stable reliability estimates in a single-level model (Furr & Bacharach, 2014, p. 348). McDonald (1999) summarized the relationship between CFA and internal consistency reliability by highlighting the role of measurement error in both CFA and estimates of reliability. Both CFA and reliability assess the amount of error variance found in a given data set. The literature on parameter estimation referring to CFA models (Anderson & Gerbing, 1984; Bearden, Sharma, & Teel, 1982; Boomsa, 1983; B. Muthén, 1983; and others) provided definitive answers regarding the appropriate number of items to improve model fit. Since the number of items is a factor known to affect reliability estimates, as discussed in detail previously, I examined the CFA literature to demonstrate the large discrepancy in recommendations made by researchers on the appropriate number of items to include in a single-level unidimensional assessment of attitudes and beliefs. Researchers proposed a minimum of four to 50 items, leaving applied researchers, educators, and clinicians to make a best guess, based on how well each item measured the latent trait of interest.

An example of CFA studies regarding the appropriate number of items to use follows: Bearden et al. (1982) conducted a Monte Carlo simulation where three items per latent trait were studied using varying sample sizes (25, 50, 100, 500, 1,000, 2,500, 5,000, and 10,000) with uninterpretable results at the smaller sample sizes. The number of items (three) did not appear to affect relative bias the parameter estimates in models with sample sizes > 500. Anderson and Gerbing (1984) conducted a simulation study where

they used five different sample sizes (50, 75, 100, 150, 300) and three levels of number of items (2, 3, 4) per latent trait to assess fit indices in a CFA. Their results suggest that models with two items are insufficient for convergence but models with three or four items per latent trait and a sample size of > 100 provided unrelative biased estimates of fit.

B. Muthén (1983) assessed the functionality of dichotomous and polytomous response choices by providing an example of a model with four items developed to measure “neurotic illness” which he found to be an appropriate number of items. Boomsa (1983) conducted a Monte Carlo simulation study in which he was examining the effects of non-normality on simple factor structures comprised of six to ten items with a varying number of response variables. Using sample size $N = 400$ with 300 replications, Boomsa (1983) found little to no relative or absolute bias in parameter estimates in the multivariate normal distribution; however, found that the model overestimated the parameters in cases of large skew (> 2). Jöreskog and Sörbom (1986) referenced a simulation study examining Likert scale items (5 response choices) and dichotomously scored items (2 response choices) in a skewed distribution and considered 5 items for measuring a single latent trait to be a small number of items and 15 items for measuring a single latent trait to be a medium number of items. MacCallum, Browne, and Sugawara (1996) suggested the number of items be related to power and effect size.

Kahn (2014) recommended between 10 and 50 items in Rasch IRT with 30 being the average number of items and in a simulation study using CFA, Jackson (2003) fixed sample sizes ($N = 50$, 100, 200. 400, and 800) and number of items (4, 5 6, 7, and 20) per latent trait to assess the N:q hypothesis (the ratio of sample size to number of estimated

parameters) using a SEM and found the 20:1 ratio to be the most appropriate which supports J. B. Kline's (1999) assertion of 10:1 or 20:1. Ziegler, Poropat, and Mell (2014) advised that as few as four items can accurately measure one latent trait within the framework of CFA, while Shevlin, Miles, Davies, and Walker (2000) recommended six items and Beeckman et al. (2010) suggested as many as 26 items should be used to measure one construct. Finally, Ulf and Lehmann (2015) suggested that a more critical issue than the number of items appropriate to measure a latent trait of interest is the "impact of an item scale on the respondent" (p. 259). Ulf and Lehmann's view is that participants look unfavorably upon assessment with multiple items and develop what they refer to as a "response style" (p. 259). The argument presented by Ulf and Lehmann posited that test taking fatigue kicks in and respondents may have a specific fallback pattern to answering items such as circling neutral for every response or using category extremes such as never and always which increases error.

The aforementioned studies supported the need for the number of items appropriate to measure one latent trait of interest in a latent variable model such as CFA or IRT to be based on *up to* three specific criteria:

1. Determine how well the items correlate with the latent trait of interest and choose items with higher correlations to include in the model.

2. A ratio of N: q of at least 10:1 and more appropriately 20:1. For example, if one has a sample size of 400 then 40 parameters, representing 20 items (for example, the mean and standard deviation of each item) would be appropriate at the 10:1 ratio where at the 20:1 ratio, a sample size of 800 would be needed to estimate 40 parameters.

3. The theoretical foundation of each item and the reasoning of how and why it will accurately measure the latent trait of interest. For example, the appropriate number of items is conditioned on how the construct is defined. A simple well-defined construct could be measured by as few as four items while a more complex construct such as IQ would likely require more items.

Finally, the debate on the number of items needed to appropriately measure the latent trait of interest in a single-level model will continue because there is no general "rule of thumb" a researcher can easily access since a construct can be narrowly defined (needing a small number of items to assess) or broadly defined (requiring a large number of items to assess). There are, however, several methods to determine the appropriate number of items based on the target population characteristics, sample size, theoretical foundations of the latent trait of interest (e.g., the definition of depression and the theory behind it), the distribution of the data, item correlations and covariance, the number of response categories, and the power and effect size a researcher wishes to achieve.

The good news is that there is some agreement in the literature regarding parameter estimation, to include reliability estimation that the number of items should range between four and 50. Based on these studies and to contain the focus of this dissertation to levels of relative bias in polychoric ordinal α and reliability estimates with both normal and non-normal distributions, the number of items in this dissertation was held to 10.

Number of Response Choices

Another factor affecting reliability estimates in a single-level model is the number of response choices. Traditionally, the 5-point Likert scale (Likert, 1932), described in

detail previously, was used in tests of attitudes. The advantages of the 5-point Likert scale, with response options ranging from strongly disagree to strongly agree, include the fact they are easily quantifiable, do not require the respondent to take a firm stand on the measured latent trait, and accommodate neutral or undecided attitudes (Johns, 2010). The disadvantages of the 5-point Likert scale include the fact a respondent's attitudes rarely fall neatly on a continuum, the continuum itself is flawed in that the distances between strongly disagree and disagree, and agree and strongly agree, are not equal, respondents often do not like to choose the extreme of one category or another (strongly disagree; strongly agree), and the true attitudes of the respondents can only be estimated and are never known. Other polytomous scales have been developed in an attempt to more precisely measure a respondent's true attitude and include modified Likert scales such as scales providing three, four, six, seven, nine, and ten choices. Nunnally (1978) found reliability to be a monotonically increasing function of the number of response choices offered and that reliability estimates accelerated up to seven response choices and evened out after eleven response choices. Several previous studies focused on how the number of response choices affect reliability. Recommendations regarding the optimal number of response options have ranged from two to three (Matell & Jacoby, 1971), six to seven (Ko, 1994), and 7 to 10 (Preston & Colman, 2000), while Aiken (1983) posited that the number of response choices does not affect Cronbach's α. The inconsistency in these recommendations leaves researchers confused when it comes to selecting an appropriate number of response choices to include. Finney and DiStefano (2006) explained that the appropriate number of response choices depends on the underlying distribution of data. For a normal or approximately normal distribution of data, "using ML estimation

techniques does not result in severe levels of relative bias [in parameter estimation] as long as the minimum number of responses is five or more" (p. 277). If the underlying distribution is severely non-normal, as the number of response choices decrease (from five or more), the greater the amount of attenuation in the parameter estimates. In other words, standard errors will increase as the number of response choices decrease causing relative biased parameter estimates. In addition to examining CTT reliability literature, to gain clarity on the issue of the appropriate number of response choices, and since latent variable models are related to internal consistency reliability, additional literature reviewed includes studies regarding how the number of response choices affect factorial validity in CFA (DiStefano, 2002; Dolan, 1994; S. B. Green et al., 1997; Hutchinson & Olmos, 1998; Maydeu-Olivares & Coffman, 2006). Lozano, García-Cueto, & Muñiz (2008) explained that even though CFA is the focus for the researchers above, their study examined the impact of varying levels of categorical response scales on evidence of validity. Categorization implies a greater loss of information over continuous data and consequently "a greater attenuation of the relationships between items" (Finney & DiStefano, 2006, p. 73). The following review of the literature regarding the number of appropriate response choices centers on how reliability estimates in single-level models are affected.

A sample of the studies that have been conducted to assess how the number of response categories affect Cronbach's α reliability estimates used empirical and/or simulated data (Aiken, 1983; Bandelos & Enders, 1996; Lozano et al., 2008; Lissitz & Green, 1975; Weng, 2004). Lissitz and Green (1975) conducted a Monte Carlo simulation to determine the relationship between reliability and the number of response

choices. They generated multivariate normal data to represent a 10-item instrument with N = 50, 100 replications per cell, six levels of response options (2, 3, 4, 5, 7, 9, 14), and three levels of item covariance (.20, .50, and .80), to assess how response choices affected estimates of reliability. Lissitz and Green found that at each of the three levels of covariance, reliability increased at 2, 3, 4, and 5 response choices and then leveled off. Lissitz and Green recommended a minimum of five response choices. Aiken (1983) conducted a study regarding whether the number of response choices affected reliability estimates. Using a 10-item teacher evaluation instrument originally developed with a 5-point Likert scale, Aiken recruited 627 participants and administered the 10-item teacher evaluation instrument with only the number of response choices changed. As expected, he found that as the number of response choices increased, the item variance increased; however, reliability coefficients remained constant. Aiken concluded that "efforts to increase the spread of responses by employing a greater number of response categories will not necessarily improve scale reliability" (p. 401).

Lozano et al. (2008) conducted a Monte Carlo simulation to assess the effects of varying data characteristics on Cronbach's α. They generated responses to 30 hypothetical items measuring one latent trait of interest and following the normal distribution. Their data included eight levels of inter-item correlations (.2, .3, .4, .5, .6, .7, .8, .9), four levels of sample sizes (10, 100, 200, and 500), and eight levels of response categories (2, 3, 4, 5, 6, 7, 8, 9). In total, 256 (8 X 4 X 8) conditions were simulated. The results showed that as response choices increased, reliability increased. The only exception was between two and three response choices, where no discernable differences were found until N = 500. These findings differ from Lissitz and Green (1975) and Aiken

(1983). For example, Lissitz and Green found differences in reliability coefficients between two and three response choices with $n = 50$ while Aiken (1983) found no evidence of reliability increasing as a function of the number of response choices. Lozano et al. (2008) concluded that (a) using only three categories was inadvisable because the majority of responses are centered at neutral and the reduction in variability affects all statistics, including reliability coefficients, and (b) when taking into account inter-item correlations and sample size, an appropriate number of response choices is four.

To address the inconsistencies in study findings outlined above, Weng (2004) recruited 1,247 participants to complete two subscales (concern for others: CO and the determination scale: DE) of the Teacher Attitude Test. The CO and DE scales were developed as unidimensional scales using a five-point Likert scale. Weng combined the CO and DE into one test with varying numbers of response choices (4, 5, 6, 7, 8, 9). As expected, the results showed that as the number of response choices increased, the means and standard deviations of both scales increased. To test for the effects of the number of response choices on coefficient alpha *the k-sample significance test* (using a χ^2 distribution) was used with a null hypothesis of equal reliability. Six conditions (4, 5, 6, 7, 8, and 9) were distributed as a χ^2 with 5 degrees of freedom and Weng found that the reliability estimates for the CO scale increased as the number of response choices increased but the reliability estimates for the DE scale did not vary with the number of response choices. Weng explained that reliability coefficient alpha was less affected by the number of response choices when the items were more homogenous (more highly correlated) and was not affected when individual variation was large. Weng concluded

that a minimum of five response choices is more appropriate when it is believed item homogeneity is high and/or individual variance is large.

Bandelos and Enders (1996) conducted a Monte Carlo simulation to determine how non-normal data affected reliability estimates Generating one normal distribution (skew = 0, kurtosis = 0), two skewed distributions (skew = 1.75, kurtosis = 1.0 and skew = 2.25 and kurtosis = 7.0), one platykurtic distribution (skew = .25, kurtosis = -1.0) and one symmetric and leptokurtic distribution (skew = 0, kurtosis = 3) representing a 10-item instrument with N=100 and three levels of inter-item correlation (.25, .5, and .75), three levels of discrete distribution shapes (item scores < 33 = 1, scores between 34 and 67 = 2, and scores > 67 = 3), and five levels of response choices (3, 5, 7, 9, 11). In total, 225 cells (5 distributions X 3 inter-item correlations X 5 levels of response choices X 3 discrete distribution shapes) were assessed and average Cronbach's α were computed for each cell design. Medium (d = .06) and large (d = .14) Cohen's d effect size estimates were used as the criteria for significance. Their results provide evidence that (a) as inter-item correlation increases, reliability increases, (b) reliability coefficients increased as a function of the number of response choices up to five categories and then stabilized, and (c) the underlying distributional shape (e.g., normal, skewed, platykurtic, leptokurtic) severely affected reliability coefficients, meaning that when data are normal or approximately normal, reliability coefficients remained stable when compared to non-normal distributions. Reliability coefficients were more likely to decrease when the underlying distributional shape and observed distributional shape (uniform, skewed, and normal) were most dissimilar. Bandelos and Enders (1996) concluded that the number of response choices should be five or more and the underlying shape of the distribution is

just as important as the inter-item correlations. This ties in nicely to the consideration of the type of underlying distribution in the next section of this dissertation. Lastly, the Bandelos and Enders (1996) study most closely relates to the goals of this dissertation. Therefore, to contain the focus of this dissertation and follow the results of Bandelos and Enders, the number of response choices was held at five.

Types of Distribution: Normal vs. Non-Normal Data

Almost all studies examining the impact of varying data characteristics on reliability coefficients used normal or approximately normal distributions. Bandelos and Enders (1996), and Sheng and Sheng (2012) generated data sets representing both normal and non-normal data and compared the results of varying data characteristics to provide guidance to applied researchers facing real world data scenarios. Enders' (2008) results show that the more non-normal the data, the lower Cronbach's α became It is important to note, however, that these effects were mediated by the magnitude of the inter-item correlations which is supported by Bandelos and Enders (1996). Therefore, when the inter-item correlations were high (.75), the shape of the distribution had less effect on the reliability coefficients. Sheng and Sheng generated data with four levels of sample size (N = 30, 50, 100, 1,000), three levels of number of items (5, 10, 30) assuming tau equivalence, three levels of Cronbach's α (.3, .6, .8), and two levels of true and error score distributions: symmetric, and non-symmetric. Table 5shows the three levels of symmetric and non-symmetric distributions assessed.

Table 5

Sheng and Sheng's Three Levels of Distribution Conditions

Skew	Kurtosis	Distribution
0	0	Normal
0	-1.35	Symmetric platykurtic
0	25	Symmetric leptokurtic
0.96	0.13	Non-symmetric
0.48	0.92	Non-symmetric platykurtic
2.5	25	Non-symmetric leptokurtic

Note. Adapted from Sheng and Sheng (2012).

A total of 432 conditions (4 sample sizes X 3 test lengths X 3 levels of reliability X 6 distributions X 2 true and error score distributions) were included in Sheng and Sheng's (2012) simulation study. Each condition involved 100,000 replications where Cronbach's α was estimated for the simulated test scores. Sheng and Sheng considered the five non-normal distributions containing α estimates as random samples from the sampling distribution α which they called *distribution α.* The final distribution was a normal distribution and the sample reliability estimates obtained were called *reliability α*, which Sheng and Sheng compared to the five distribution αs. The criterion for significance (testing the hypothesis that reliability α = distribution α) is that if the observed mean of distribution(s) α = reliability α, then distribution α is not significant (unrelative biased). If the observed mean of distribution(s) α ≠ reliability α, then distribution α is significant (relative biased) and either positively or negatively relative biased based on whether it is larger or smaller than reliability α. The results showed that

(a) skewed or platykurtic distributions did not affect Cronbach's α; however, (b) both symmetric and non-symmetric distributions with high kurtosis resulted in smaller mean alphas and larger standard errors equating to wider 95% confidence intervals, and (c) non-normal distributions with high positive kurtosis underestimated Cronbach's α more so than other distributions tested. These findings support Enders' (2008) findings regarding the effect of non-normal distributions on Cronbach's α. It is important to note that the aforementioned studies regarding normal and non-normal data distributions generated all items to be either normal or non-normal and did not mix item distributional characteristics, which can be found in real-world data. Therefore, the effect of mixing item distributions within one generated data set may provide interesting results important for applied researchers. One goal of this dissertation was to mix item distributions within one data set and examine levels of relative bias in reliability estimates. This is discussed in more detail in Chapter III.

Multilevel Model

These psychometric analyses presented to this point refer to single-level models not often found in applied behavioral and educational research and rely on the assumption of independence of observations. Factors affecting reliability in multilevel models are less certain (Geldhof et al., 2014) and are addressed in the following section of this dissertation.

Appropriate sampling designs place the data collected into proper context. Behavioral, educational, and social science data are not collected in a vacuum but rather are imbedded in a hierarchy of environment (Luke, 2004). For example, Luke (2004) explained that "the likelihood of developing depression is influenced by social and

environmental factors" (p. 1), health outcomes vary based on socioeconomic status, and educational outcomes vary based on a number of social and environmental conditions. Despite the importance of context, studies in the fields mentioned above often focus on single-level analysis (the hopeless individuals, specific health outcomes for individuals, attitudes of fifth graders) without accounting for the influences higher levels of context have on individual scores. Presenting data in the proper context provides a deeper understanding of how variance is influenced by higher level factors and reduces unexplained variance.

B. Muthén (1989), in supporting the results of Lord and Novick (1968), showed that if unobserved heterogeneity (e.g., due to unexplained grouping) is ignored, it can lead to inflated measurement reliability. An example of polytomously scored data in educational research would be if a researcher wanted to assess fifth graders' attitudes regarding standardized tests in Weld County. The researcher may randomly select 200 fifth graders across Weld County, give them the attitude survey, and evaluate the composite scores. In this example, the schools and classrooms from which the students were enrolled were not taken into account and yet may help elucidate the unexplained variance in the attitude scores of the fifth graders. In other words, a more appropriate sampling design would be to take into account the nested aspect of the data (e.g., the hierarchy of the environment) which will allow the researcher to assess any effects school or classroom may have on the attitude score results. While multilevel data structures can have more than two levels and include time as a repeated measures variable, the focus for most researchers, and the focus of this dissertation, is the two-level cross-sectional model (Bryk & Raudenbush, 1992, Geldhof et al., 2013; Raudenbush & Bryk, 2002, Raykov &

Penev, 2010), and for the sake of clarity, levels of analysis will be referred to as individual (level-1) and group levels (level-2) with a single-level model focusing on individual effects and a two-level model examining individual effects at level-1 and group effects at level-2 (Geldhof et al., 2014; Preacher et al., 2012; Raudenbush & Bryk, 2002).

Sample Size in a Multilevel Model

The sample size recommendations within the CTT and Rasch IRT frameworks described above are for single-level models only and do not take into account multilevel sampling designs most often found in educational, psychological, and social research. Snijders (2005) reviewed sample size requirements in multilevel modeling. While not directly discussing sample size effects on reliability coefficients, Snijders explained that to accurately estimate model parameters such as mean, variance, and effect size of a level-one variable, "the sample size at the highest level is the main limiting characteristic of the design…for testing the effect of a level-two variable it is the level two sample size . . . [that is the most important]" (p. 1571). For example, in a two-level model where patients are nested within clinics or students are nested within schools, the lowest level would be the patients or students and the second (or higher) level would be the clinics or schools. Expressly, the sample size at level-two (clinics or schools) is more critical in determining sample size requirements than the sample size at level-one (patients or students). Snijders provided a formula for computing appropriate sample sizes based on design effects (*deff*) as represented by Equation 21:

$$deff = \frac{squared\ stadard\ error\ under\ the\ design}{squared\ standard\ errir\ under\ the\ standard\ design} \quad (21)$$

where the “standard design” is defined as a multilevel design where the sample size at level-one is the same as other levels in the model. He explained that since variances of parameters are inversely proportional to sample sizes, multiplying *deff* by a sample size collected in a single-level model using simple random sampling techniques would provide an appropriate sample size for the level of interest in a multilevel design. B. Muthén and Satorra (1995) found that a *deff* < 2 using single-level analysis of multilevel data had a negligible effect on parameter estimates; therefore, Maas and Hox (2005) used a *deff > 2* to determine appropriate sample sizes for multilevel models.

Building on the work of Busing (1993) and Van der Leeden and Busing (1994), Maas and Hox (2005) studied the effects of sample size in parameter estimation on such entities as regression coefficients and variances. Using predetermined intraclass correlations (ICC = .1, .2, .3) at both level-1 (individual) and level-2 (group), Maas and Hox simulated data to represent the number of groups (NG = 30. 50, 100) and group size (GS = 5, 30, 50) based on simulations by Van der Leeden et al. (1997). For each of the 27 conditions (3 ICCs X 3 NG X 3 GS) it was assumed the data were normally distributed and the explanatory variables (individual and group levels) were fixed. The results showed that the variance components were stable across all data conditions at level-1; however, the standard errors at level-2 were underestimated when NG < 100. Further analysis provided evidence that when NG = 50 the standard errors were not underestimated as frequently as when NG = 30. They concluded that NG = 50 at level-2 was acceptable for multilevel modeling. These results differ from the results found by Busing (1993) and Van der Leeden and Busing (1994) where the minimum number of groups (level-2) recommended was > 100. Bell, Ferron, and Kromrey (2008) conducted a

Monte Carlo simulation to determine the effect of small sample sizes at level-1 and level-2 of a multilevel model. Their data conditions for level-1 included small sample sizes (average = 10, range 5-15) and large sample sizes (average = 50, range = 25-75) and five items. For level-2, sample sizes were *N* = 50, 100, 200, and 500 with four items. In all, 5,760 conditions were tested and their results suggested stable parameter estimates at both levels of the model, with the exception of level-2 *N* =50, where the confidence intervals were found to be less accurate. These findings support previous simulation studies.

Number of Items, Response Choices, and Distributions

A review of the reliability estimation in multilevel literature shows that there is insufficient advice provided regarding the appropriate number of items and response choices needed to reduce bias in estimates of reliability in multilevel models. Geldhof et al. (2014) encourages further research into varying and assessing these data characteristics in multilevel models. In this dissertation, I investigated the effect of sample size and normal, non-normal, and mixed data distributions while holding the number of items at ten and the number of response choices at five, which is outlined in Chapter III.

Building a Two-Level Model

Two common approaches to multilevel modeling are multilevel regression models and multilevel factor analysis models. The multilevel regression model, also known as hierarchical linear modeling (HLM; Bryk & Raudenbush, 1992; Raudenbush & Bryk, 2002), random coefficient models (Rosenburg, 1973), or covariance component models (Dempster, Rubin, & Tsutakawa, 1981), permit the partitioning of variance, critical for

analyzing hierarchically nested data such as the data in the previous example where students are nested within classrooms. Researchers may specify a level-1 model where the parameters of the model illustrate a linear relationship between level-1 units. These level-1 parameters "are then viewed as varying across level-2 units as a function of level-2 characteristics" (Raudenbush, 1993, p. 462). Consider the previous example of a two-level model from above where student attitude scores are at level-1 and classroom effects are at level-2. The model for these data can be conceptualized as a multilevel regression model or a multilevel factor analysis model depending on the parameter estimations of interest (Maydeu-Olivares & Coffman, 2006). In this dissertation, since internal consistency reliability estimates are of interest, the multilevel factor analysis approach is discussed. Within the multilevel factor analysis model, the aforementioned example of a two-level model can then be considered as a random intercept model because the intercepts can be treated as a random effects and the items are used to explain the latent trait of interest (factor) in the model (Maydeu-Olivares & Coffman, 2006).

Whether the reliability coefficient is anchored in CTT or assessed within the framework of Rasch IRT will affect how reliability is estimated in a multilevel model. In this dissertation, Cronbach's α and polychoric ordinal α will be examined using a multilevel factor analysis model. Reliability estimates within the Rasch IRT framework will be assessed using person and item parameters, which are discussed in detail in the section below.

Using Confirmatory Factor Analysis to Estimate Reliability in a Two-Level Model

Single and two-level models were reviewed previously in regard to selecting the appropriate sample sizes. The discussion included several references to the effect of single and two-level models on reliability coefficients which are only highlighted, and not rewritten, in this section. Recall, B. Muthén and Sattora (1995) found that single and two-level models did not affect parameter estimation while Maas and Hox (2005) found smaller standard errors of measurement (SEM) in a two-level model than a single-level model, which could lead to overestimation of the reliability coefficient in level-2, since SEM is related to reliability (e.g., as the reliability coefficient increases, the SEM decreases since higher reliability means lower error; Biemer, Christ, & Wiesen, 2009). Recall that Cronbach's α is the ratio of true score variance to the total score variance and the goal of reliability analysis is to obtain unrelative biased estimates of measurement error. Further, Snijders and Bosker (1999) suggested that since multilevel sampling confounds within-group variance and between-group variance, it may lead to relative biased reliability estimates since the assumption of independent residuals is violated. Few studies have addressed the impact of two-level models on reliability estimates even though the need to account for variability in multilevel models has been described in and well established by the literature (B. O. Muthén, 1994; B. Muthén & Asparouhov, 2011; Raykov & Penev, 2010; Snijders & Bosker, 1999). For example, B. O. Muthén (1994) detailed the "perspective of varying parameters" (p. 377) as they related to multilevel structures. He explained that data in multilevel models are often obtained via cluster sampling techniques and studied by comparing the ratio of the variance of the estimator

under cluster sampling conditions to the variance under simple random sampling conditions such as those found in single-level models.

B. Muthén and Asparouhov (2011) described several types of statistical analyses relevant for clustered data and recommended a two-level regression analysis, two-level path analysis, two-level EFA, two-level latent variable modeling, multilevel factor analysis, or a two-part growth model to examine both within-subjects and between-subjects' variability. To provide guidance to applied researchers on the effects of single- and two-level models on reliability coefficients, Raykov and Penev (2010) and Geldhof et al. (2014) conducted hypothesis tests to assess traditional and nontraditional reliability coefficients in a multilevel model using continuous data. Raykov and Penev (2010), Geldhof et al. (2014), Black et al (2015), Yang, Beitra, and McCaffrey (2015), Huang and Cornell (2016), and T. A. Brown (2015) suggested using some form of latent variable modeling (LVM) techniques to estimate reliability in multilevel models. Within each study, these researchers were interested in modeling the outcome (dependent) variable, y_i, as a function of lower and higher sample levels (e.g., individuals and groups). I found multilevel factor analysis to be more relevant to my dissertation than the group means approach presented by Raykov and Penev or the composite reliability approach of Yang et al; therefore, further discussion details the research conducted by Geldhof et al. (2014), Huang and Cornell (2016), and T. A. Brown (2015) and their corresponding results.

Multilevel Confirmatory Factor Analysis

Geldhof at al. (2014) conducted a simulation study to provide recommendations when assessing reliability in a multilevel model. Stating that the basic CFA model can "be elaborated in various ways" (p. 76), the researchers restricted their focus to continuously

scored items. Since polytomously scored items have more often been found in an assessment of attitudes, this data characteristic is at the center of my dissertation and the MCFA model for polytomously scored items is highlighted in section describing the Huang and Cornell (2016) study and described in detail in Chapter III. While Geldhof et al. (2014) recognized the contribution made by Raykov and Penev (2010) regarding LVM, they believed that "the reliability of group means as estimates of the distribution of means in a population is different than measurement reliability" (p. 74). Instead, Geldhof et al. suggested that reliability estimated at each level of a two-level model within the framework of CFA (known as multilevel confirmatory factor analysis or MCFA) is a better approach since general reliability coefficients may be relative biased when the assumption of independent residuals is violated. Multilevel sampling will result in hierarchically structured data, as mentioned previously, "making the residuals dependent in the presence of between-cluster variation" (Geldhof et al, 2014, p. 72). Most behavioral, educational, and social science researchers who use multilevel modeling to account for the variance at each level of analysis tend to then report Cronbach's α as a measure of reliability, which implies a single-level data structure since it uses a scale's total variability rather than measuring reliability at each level of analysis (T. A. Brown, 2015; Cronbach, 1951; Geldhof et al., 2014). According to Geldhof et al., single-level reliability estimates using CFA summarize the factor loading matrix into an easily interpretable result. Cronbach's α can then be used to estimate reliability. Geldhof et al. (2014) posited that this approach can be extended to a two-level model by "specifying fully saturated indicator covariance matrices in both levels of a [multilevel confirmatory factor analysis] MCFA" (p. 76) and estimating Cronbach's α at the within-level and the

between-level. Using data simulated to be continuous, Geldhof et al. explained that the MCFA model they chose provided a similar form of decomposing variance as found in generalizability theory, where an individual's observed score on an item can be decomposed into four distinct parts as represented in Equation 22:

$$\underbrace{Y_{ik} = T_{wi} + E_{wi}}_{\text{within-cluster}} + \underbrace{T_{bk} + E_{bk}}_{\text{between cluster}} \quad (22)$$

Where T_{wi} is the true deviation from the cluster average true score, E_{wi} is the within cluster error, Tbk is the individual's cluster average true score, and E_{bk} is between cluster error. Using this model, true score variance can be acquired at each level. Reliability at the within level is the ratio of within-cluster true score variance to total within-cluster variance and reliability at the between level is the ratio of between-cluster true score variance to total between-cluster variance. Geldhof et al. (2014) further explained that since the between-cluster reliability is represented in a scale, it "does not necessarily represent the reliability of group-level composites" (p. 75). Therefore, between-cluster reliability is different from ICC which is the ratio of between-cluster variance to its total variability across both levels. Geldhof et al. further explains that the idea of more than one error term is contrary to CTT but concludes that a multilevel model, by its very nature, requires the assessment of observed score and error variances at each level and therefore, the MCFA approach is appropriate because, unlike generalizability theory, "MCFA decomposes observed scores [(X)] into components related to each individual's cluster average true score (T)" (Geldhof et al., 2014, p. 75). In other words, MCFA follows CTT requirements by combining item specific variance, between-cluster

differences, the interactions among them, and variance due to nonsystematic error into one residual term. The MFCA model provided by Geldhof et al. is given as a special case of B. Muthén and Asparouhov's (2009) model by a set of three equations (Equations 23, 24, and 25; Geldhof et al., 2014) which assume a continuous scale.

$$Y_{ij} - \Lambda_j \eta_{ijj} \tag{23}$$

$$\eta_{ij} = \alpha_j + \beta_j \eta_{ij} + \zeta_{ij} \tag{24}$$

$$\eta_i = \mu_{jj}\, \beta \eta_{ijj} + \zeta_j \tag{25}$$

where *subscript i* represents level- 1 (individual) and *subscript j* represents level-2 units (groups). Y_{ij} is a vector of p measured variables; $\Lambda_j = \Lambda = [Ip\ 0p\_m\ Ip\ 0p\_m]$ is a (p X ($2p$ + $2m$) factor loading matrix linking Yij to p latent parts at both the within- and between-cluster levels, and m common factors at both levels; αj is a vector of length ($2p$ X $2m$) containing p latent within-cluster parts, m within-cluster common factors, p latent between-cluster parts, and m between-cluster common factors; ηj is a vector of length ($2p$ X $2m$) that contains the p item intercepts and m between-cluster common factors; Bj is a ($2p$ X $2m$) X ($2p$ X $2m$) matrix containing within-cluster factor loadings; ηj (r X 1) contains all of the j-subscripted random coefficients from αj and Bj, including the between-cluster common factors; μ (r X 1) contains means of those coefficients and the item intercepts; β (r X r) contains between-cluster factor loadings; ζij contains unique factors and common factor residuals for the within-cluster model; and ζj (r X 1) contains unique factors and common factor residuals for the between-cluster model. Separate within- and between-group α can be obtained by applying Equation 25 to the

between- and within-group results. Equation 26 represents α as a function of the average inter-item covariance within a scale (mean σij), the variance of the scale (σ2x) and the number of items included in the scale (n) reflected by Equation 26 (Cronbach, 1951):

$$\alpha = n^2(mean\ \sigma_{ij})/(\sigma^2_X) \quad (26)$$

Note that α can be estimated in MCFA by "specifying a fully saturated covariance structure model that has no latent variables" (Geldhof et al., 2014, p. 73). In this way, the MCFA method leads to observed scores (*Yij*) encompassing both true score and error variance at both within-cluster and between-cluster levels denoted by subscript *ij*. Revisiting the previous example of assessing fifth grade attitude scores in a two-level model, this means that the MCFA approach will permit attitude score variances and covariance to vary at level-1 and level-2.

Focusing only on the aspect of their study where data were generated to reflect a two-level model and estimating Cronbach's α, Geldhof et al. (2014) hypothesized that (a) ignoring the hierarchical data structure will make reliability estimates difficult to interpret unless reliability is equal across both levels; (b) as the ICCs decrease, the reliability estimates in a single-level model will roughly reflect the within-level (level-1) reliability estimates and as the ICCs increase, the reliability estimates will roughly reflect between-group (level-2) reliability estimates; (c) MCFA may fail to reproduce an underlying factor structure when item reliabilities are low, especially in the face of low sample sizes; and (d) using a fully saturated two-level model to estimate alpha, no convergence problems will exist. Geldhof et al. (2014) examined reliability estimates for a six-item congeneric measurement model (scale). A congeneric measurement model means that each item is

related to only one latent trait (unidimensional) and all covariation between items is a consequence of the relationships between items and the latent trait of interest. They simulated continuous data to represent three conditions of observations per cluster (number of individuals: 2, 15, 30), three conditions of the number of clusters (groups: 50, 100, 200), four conditions of ICCs (.05, .25, .50, .75), both low and high reliability conditions (α = .30 and .85) and three conditions of factor loadings set to .8, .7, and .6 for both level-1 and level-2. First, the researchers calculated the bias of Cronbach's α in the single-level model under a cross-section of all of the data conditions, and then they compared these results to reliability estimates obtained from simulating a multilevel model with either low or high ICCs (.05 and .50), the total number of observations (200 clusters with 30 observations each and 100 clusters with two observations each), and conditions where reliability was high at both levels, neither level, only within level-one, and only between level-two clusters. Finally, assessing the root mean squared error of approximation (RMSEA), the comparative fit index (CFI), and the Tucker-Lewis index (TLI) fit statistics, and 95% confidence intervals, their results show that within-level alpha was never biased more than 10% and considered acceptable in all conditions; however, between-level alpha was negatively biased for small clusters (50, 100) when the ICCs were low (e.g., .05, .25) and within-level reliability was low (e.g., .30). Geldhof et al. (2014) concluded that reliability estimation should be level-specific when working with multilevel data. Furthermore, within-level reliability estimates and between cluster reliability estimates were acceptable under all conditions except when the number of observations = 2 and the ICCs were low, in which case they were underestimated.

The random intercept model in MCFA is described by Geldhof et al. (2014) and T. A. Brown (2015). This model allows intercepts to vary. Reconsidering the fifth graders' attitudes outlined previously, the goal would be to specify an MCFA where a single factor of attitude is specified at both within and between levels to explain the variability of the items on the survey. T. A. Brown (2015) clarified the process of building an appropriate MCFA model by recommending multiple steps to estimating reliability in an MCFA. First, one should examine the ICCs of the items, as they refer to the proportion of variance in the items due to the clusters. If the ICCs are < .05 then a multilevel model may not be necessary when estimating reliability. Second, "[specify] a CFA model at the within-level leaving it unstructured at the between-level" (T. A. Brown, T., 2015, p. 421). Third, if an appropriate measurement model exists, "examine the between-level factor structure in a two-level model with the within-level structure fully specified" (T. A. Brown, 2015, p. 421). Equations 27, 28, and 29 represent the within-and between-level model (T. A. Brown, 2015, p. 421) modified from Raudenbush and Bryk (2002):

$$Y = \Lambda_w \eta_w + \varepsilon_w \qquad \text{(within)} \tag{27}$$

$$\mu_B = \mu + \Lambda_B \eta_B + \varepsilon_B \qquad \text{(between)} \tag{28}$$

which can be combined as:

$$Y = \mu + \Lambda_w \eta_w + \Lambda_B \eta_B + \varepsilon_{B+} + \varepsilon_w \tag{29}$$

Where μ is the vector of between-level means; Λ_w is the within-level factor loading matrix; η_w is the within-level factor; and ε_w is the item residual variance within-levels; Λ_B is the between-level factor loading matrix; η_B is the between-level factor, and

ε_B is the item residual variance between levels. The factor loading matrices ($\Lambda_{w,}$ Λ_B) and cluster level means (μ) are fixed effects while μ_B refers to the random intercepts of the Y variable. Note that the MCFA model given as a special case of B. Muthén and Asparouhov's (2009) model and the MCFA model specified above are both fully specified models using continuous data. The main differences between Geldhof et al.'s (2014) model and T A. Brown's (2015) model are (a) the way in which each is expressed, with Brown's model providing more clarity and (b) how factor loadings are handled, with Geldhof et al. fixing all factor loadings to 1.0 and T. A. Brown fixing only the first factor loading to 1.0 and allowing the remaining factor loadings to be freely estimated. Figure 3 illustrates the path diagram for the previously discussed example of a two-level model where student attitude scores (considered continuous in this example) are at level-1 and classrooms are at level-2. In the following example, there are four hypothetical (observed) items on the attitude scale in Figure 3 (Y1, Y2, Y3, Y4) which represent the within-level measurement and are depicted by squares. These four items are then considered four latent variables at the between-level measurement and depicted by small circles. The two large circles in the path diagram are the attitude factor at both the within level and between level. The single between-level factor *Attitude between* is specified to account for the variation and covariation among these random intercepts. Attitude within is the within-level attitude factor with four items and Attitude between is the between-level factor where the four items in level-1 are considered latent variables in level-2: Brown provided M*plus* commands to build the a two-level MCFA which is included in Chapter III and used to estimate within-and between-level reliability.

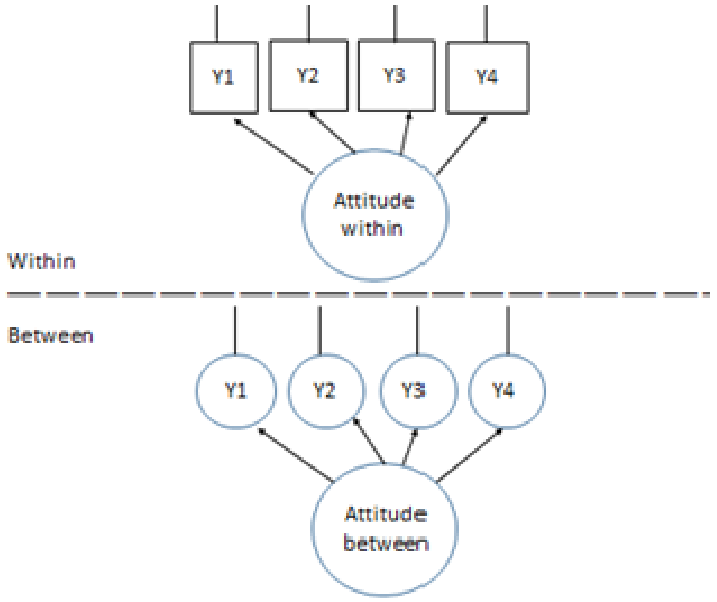

Figure 3. A hypothetical path diagram of a Multilevel Confirmatory Factor Analysis model of attitude.

Finally, Geldhof et al. (2014), Black et al. (2015), and T. A. Brown (2015) provided the same recommendations regarding model fit and convergence rates. In other words, there was consensus among the researchers that good model fit (e.g., CFI > .95, TLI > .95, RMSEA < .05) must be present in order to provide meaningful interpretations of reliability at the within-and between-levels of analysis. Additionally, it is important to note that Geldhof et al. (2014), Black et al. (2015), and T. A. Brown (2015) considered only continuous data in their research. Their MCFA models all found stable within-level Cronbach's α but biased estimates of Cronbach's α at the between-level. Geldhof et al. reported negative bias at the between-level Cronbach's α when ICCs were low (< .05), Black et al. (2015) reported the between-level Cronbach's α to be underestimated

regardless of ICCs, and Brown did not recommend Cronbach's α to estimate between-level reliability.

Each study estimating reliability in a multilevel model offered important ideas for the methods to employ in this dissertation; however, since my focus was specifically on polytomous data, by extending the MCFA model to polytomous data, Huang and Cornell (2016) provided valuable methodological techniques critical for this dissertation. Huang and Cornell conducted an MCFA to examine the factor structure of the Positive Values Scale (PVS). Their participants included 39,364 seventh-and eighth-grade students from 423 schools. Each school randomly chose a sample of students to whom to administer the PVS. The PVS is a unidimensional measure of positive values and is comprised of nine items with six categorical response choices for each item, from not important to extremely important. Huang and Cornell found good model fit with RMSEA = .04, CFI=.98, and TLI =.97 and low ICCs across all items (<.05). At level-2, item 2 resulted in a small negative residual variance and was therefore fixed to 0 to allow for convergence of the model as recommended by Hox (2002) and Asparouhov and Muthén, (2006). Huang and Cornell then calculated Cronbach's α at level-1 (within-students level) and to account for the clustered nature of the data and following the guidelines of Dedrick and Greenbaum (2010), used the Spearman-Brown prophecy formula (see Equation 30), as a measure of reliability at level-2 (between-school level).

$$[k(ICC]/[(k-1)(ICC) + 1] \quad (30)$$

Where k is the average number of respondents per school and the ICC is the intraclass correlation of the factor that reflects the amount of variability resulting from the school-

level. Cronbach's α for level-1 was .81 and the Spearman-Brown reliability coefficient = .92 providing evidence of good reliability at both the within-and between-levels.

The focus of this dissertation was to examine the behavior of Cronbach's α along with Spearman-Brown's Prophecy formula to provide guidance to applied researchers on the most appropriate reliability coefficient(s) to use in a multilevel model. Therefore, Cronbach's α was estimated at the within-and between-levels using Geldhof et al.'s (2014) and T. A. Brown's (2015) approaches, as well as the Spearman-Brown Prophecy formula for the between-level approach found in Huang and Cornell (2016). In addition, confidence intervals for each reliability estimate are reported as recommended by Geldhof et al. (2014), Black et al. (2015), T. A. Brown (2015), and Huang and Cornell (2016). These approaches in MCFA were compared to the other reliability estimates: polychoric ordinal α within the CTT framework and person and item reliability and separation within the Rasch IRT (RSM) framework.

Assessing Reliability Within the Rasch Item Response Theory Framework

Assessing reliability in the Rasch IRT model requires examining both person and item parameters. Kamata (2001) explained that several methods are used in estimating these parameters based on multilevel data. The two methods applicable to estimating person and item parameters for this dissertation are described here. First, person parameters are **treated** as random effects rather than the classical fixed-effects model described by Rasch (1960). They can be decomposed into a linear combination of fixed and random effects. This method allows the researcher to perform analysis of person characteristics such as person reliability and separation. Second, a multiple group IRT can

be used to group individuals by a common characteristic such as classroom or school. A multiple group IRT assumes separate latent distributions for groups in estimating item parameters. Kamata proposed a three level formulation where level-one is the item level, level-two is the person level, and level-three is the group level (e.g., classrooms or schools). Building on and applying Kamata's work, Raudenbush, Johnson, and Sampson (2003) explained that Rasch IRT scaling presents person-specific standard errors of measurement as well as item-specific and whole-scale information making it a good choice to use in multilevel models. Their study was in regard to self-reported crime statistics where the respondents were nested within social settings such as schools or neighborhoods. Following Kamata, Raudenbush et al. treated person parameters as random effects. Randenbush et al. then applied the Kamata approach where level-1 includes item responses which are dependent on item difficulty and person propensity (e.g., the likelihood of committing a crime, which is interpreted as person ability with higher levels indicating higher likelihood to commit a crime). Level-2 describes variance and covariance between person propensities (e.g., abilities) within groups (e.g., schools or neighborhoods), and level-3 describes variance and covariance between groups. Reliability can then be assessed at each level by examining level-specific person and item reliability and separation. Linacre (2014) noted that reliability and separation parameters should be considered together when making decisions based on reliability of the scores. Confidence intervals are calculated to determine any reliability relative bias. By conceptualizing criminal behavior in this way, Raudenbush et al. (2003) were able to model crime indicators for both individual differences and contextual variation. In other

words, individual self-report responses on the likelihood to commit a crime and the hierarchical structure presented as social settings.

Pastor (2003) followed the recommendations of Kamata (2001) and conducted a study using Kamata's three-level IRT model. Pastor explained that Kamata's three-level model provided four distinct advantages over other multilevel modeling techniques within the Rasch IRT framework. These advantages are that hierarchically structured data can be modeled, latent traits of interest can be estimated at different levels, improved estimates of inter-item correlations and relationships between latent traits and observed variables can be calculated, and these relationships can be examined at different levels of analysis. Using the Culture Free Self-Esteem Inventories (CFSEI-3), Pastor first built an unconditional model which modeled the variation of item responses within people and used the log-odds of the probability of endorsing an item for a given person. Item effects were specified across persons so that in level-2 of the unconditional model, only variation among persons in level-1 within level-2 groups were estimated. The third level was used to model variation among groups by using the parameters estimated for each group where item effects were fixed across groups and latent trait effects varied randomly across groups. Considering the three-level model, once Pastor determined significant variation across level-1 variables, person variables of gender and age were modeled at level-2. Finally, group variables were modeled in level-3 to assess group level variation. Pastor compared the unconditional model to the conditional model with item and group variables at levels 2 and 3 and found improved model fit. While her study did not examine person or item reliability, it does provide information on building the unconditional three-level model which she stated was based on the Rasch IRT model

(e.g., satisfied the assumption of equal discrimination across items) and is, therefore, important to include in my dissertation., Once the three-level unconditional model is specified, person and item reliability and separation can be estimated.

Chapter Summary

Observed scores on an instrument designed to measure a latent trait such as aptitude or attitude contain some element of error. Measurement theories such as CTT (Spearman, 1904) and Rasch IRT (Rasch, 1960) were established to measure not only the observed scores but the various elements of error inherent to data collected from human subjects. Reliability coefficients were developed to measure error in composite test scores and evolved to measure error at the item level. This elevated social science research beyond descriptive statistics and into the world of hypotheses testing. Each of these new measurement frameworks (e.g., CTT, CFA/MCFA, and Rasch IRT) carry assumptions often not met with data collected from human subjects. Discussions regarding the selection of the most appropriate sample size, number of response choices, and sampling design to reduce error in survey research continues. The purpose of this dissertation was to simulate observed scores under complex data conditions often found in the real world and (a) investigate error in terms of internal consistency reliability within the CTT framework and person and item reliability and separation within Rasch IRT models (e.g., RSM), (b) inform clinicians, teachers, and applied researchers about possible relative bias in reliability coefficients when more complex data structures and underlying distributions are encountered, and (c) provide a reference from which to interpret their results.

In Chapter III, the methods developed to answer the research questions in Chapter 1 are described in detail and examples of programming codes are included.

CHAPTER III

METHODS

Introduction and Research Questions

In Chapter III I discuss the methodology used to answer the four research questions posed in this dissertation. Three principle objectives were addressed in this dissertation: (a) following a call from Zumbo et al. (2007) and Gadermann et al. (2012). I examined the effect of sample size, type of underlying data distribution (normal, non-normal, or mixed), sampling design (single and two-level), and the interactions of these data conditions on estimates of polychoric ordinal α within the CTT framework; (b) following the recommendations of T. A. Brown (2015), Geldhof et al. (2014), Huang and Cornell (2016), Little (2013), and Sheng and Sheng (2012), I examined the effect of sample size and type of underlying data distribution on estimates of Cronbach's α within the CTT framework (single-level design) and MCFA framework (two-level design); and (c) examined person reliability in a two-level RSM model. Using a fully crossed design, these analyses focused on the possible consequences of these varying conditions on estimates of reliability. Monte Carlo simulation techniques were used to generate data to answer the following four research questions as they pertain to a unidimensional measure with polytomous data:

Q1 In a single-level model, to what degree do data conditions (sample size and distribution of data) affect levels of bias in reliability estimates (a comparison of Cronbach's α, polychoric ordinal α, and person reliability)?

H1 In single-level models, bias in reliability estimates will increase under the conditions of smaller sample sizes and non-normal or mixed distributions and polychoric ordinal α and person reliability will be less biased than Cronbach's α.

Q2 In a multilevel model, to what degree do data conditions (sample size and distribution of data) affect levels of bias in reliability estimates (a comparison of Cronbach's α, polychoric ordinal α, and person reliability in level-1 (within groups) and the Spearman-Brown's prophecy coefficient in level-2 (between groups)?

H2 In multilevel models, bias in reliability estimates in level-1 will increase under the conditions of smaller sample sizes and non-normal or mixed distributions and polychoric ordinal α will be less biased than Cronbach's α and person reliability. Additionally, Spearman-Brown's prophecy coefficient will be underestimated under the conditions of smaller sample size and non-normal or mixed distributions

Q3 Do standard errors and levels of bias in reliability estimates (Cronbach's α, polychoric ordinal α, and person reliability) differ when data are single-level versus when data are at level-1 of a two-level across sample size and distribution of data?

H3 When comparing the standard errors and levels of bias in reliability estimates of single-level and level-1 of two-level sampling designs, across three estimates of reliability, bias for level-1 of the two-level model will be lower than the bias found in the single-level models.

Q4 To what degree do interactions among sample size, data distribution, and sampling design (e.g., single-level and two-level) affect levels of bias in reliability estimates (Cronbach's α, polychoric ordinal α, person reliability, and Spearman-Brown's prophecy coefficient)?

H4 Interactions among sample size, data distribution, and sampling design will increase bias in reliability estimates, with the joint effect of lower sample sizes and non-normal and/or mixed distributions displaying the most bias.

The Pilot Study

Data generation and the final selection of data characteristics of this dissertation are discussed in detail in the next section. During the proposal phase of this dissertation, a pilot study was conducted on a small portion of the myriad of proposed data conditions to

(a) assess my ability to simulate multivariate normal data within the CTT framework for single-level models in R and (b) determine the data conditions used in the full study. Additionally, within the Rasch IRT framework, data representing a single-level RSM were also generated using R (see Appendices A and B, respectively) and analyzed using Winsteps. To manage these data sets, all reliability coefficients were estimated in R and exported into MS Excel, and 95.0% confidence intervals about the sample reliability estimates were calculated and relative bias were examined and trends in bias elucidated.

The pilot study included generating multilevel data; however, these data were simulated using Hierarchical Linear Modeling (HLM) techniques rather than MCFA techniques. During the proposal defense, MCFA techniques were found to be more appropriate for estimating reliability coefficients in a multilevel model. Therefore, the multilevel data conditions and subsequent results from the pilot study are not included here.

Cronbach's Alpha

Multivariate normal data were generated in R and the same seed was used in all single-level analyses. Cronbach's αs and their corresponding 95.0% confidence intervals, and person and item reliability and person and item separation indices in RSM were estimated. An example of data conditions used for the pilot study and all resulting biases are found in Table 6

Table 6

Summary of Pilot Study Data Conditions

	Single-Level Sampling Design
Distributional Characteristics	Normal [Skew = 0, Kurtosis = 0]
For CTT:	
Cronbach and polychoric ordinal α*	.70
Sample size(s):	
Individuals *(N)*	*N* = 50, 200
Number of:	
Items *(i)*	*i= 5, 10*
Response choices (rc)	rc = 4, 7
For PCM and RSM:	
person reliability*	.70
Person ability*	μ = 0, σ = 1
Item difficulty*	μ = 0, σ = 1
Sample size(s):	
Individulas *(N)*	*N* = 50, 200
Number of:	
Items *(i)*	*i = 5*
Response choices (rc)	*RC* = 4

Note. All data generated will represent a unidimensional model measured by polytomously scored items.

* indicates fixed parameters

Pilot Study Results

Single-level Cronbach's α. Table 7 represents the Cronbach's α results and corresponding 95% sample confidence intervals and standard errors for the eight single-level data sets generated. Within the eight 95% sample confidence intervals calculated from the samples drawn, the known population reliability coefficient (.70) was captured 100% of the time.

The results of the factorial ANOVA with absolute bias as the dependent variable and sample size (*n*), number of items (*i*), and number of response choices (*rc*) indicated as the random factors is in Table 8 below. No statistically significant absolute biased estimates were found under these data conditions based on all interaction and main effects having *p*-values > .05.

Table 7

Pilot Study: Cronbach's α and Corresponding 95% Confidence Intervals from the Single-Level Sampling Design and Normal Distribution

Sample size (*N*)	Number of items (i)	Number of response choices (rc)	Population Cronbach's α	Average (observed) Cronbach's α	95% Sample Confidence Intervals		SE of Sample Cronbach's α	Relative Bias	Absolute Bias	Population Cronbach α falls within the 95% Sample Cronbach's a Confidence Interval
					Lower Level	Upper Level				
50	5	4	0.70	0.6567	0.4903	0.8231	0.08	-0.24	0.04	yes
50	10	4	0.70	0.6649	0.5113	0.8185	0.08	-0.25	0.04	yes
50	5	7	0.70	0.6559	0.5069	0.8049	0.07	-0.24	0.04	yes
50	10	7	0.70	0.6559	0.5021	0.8097	0.08	-0.24	0.04	yes
200	5	4	0.70	0.6567	0.5831	0.7303	0.04	-0.24	0.04	yes
200	10	4	0.70	0.6629	0.5935	0.7323	0.03	-0.25	0.04	yes
200	5	7	0.70	0.6708	0.5974	0.7442	0.04	-0.26	0.03	yes
200	10	7	0.70	0.6773	0.6069	0.7477	0.04	-0.27	0.02	yes

Table 8

Tests of Between-Subjects Effects for Bias in Cronbach's Alpha in a Single-Level Model

Source	*F*	*p*-value
Intercept	9.44	.33
Same size (*n*)	1.000	.50
Number of items ()	1.000	.67
Number of response choices (rc)	1.000	.50
Sample size * Number of items	1.000	.50
Sample size * Number of response choices	9.000	.20
Number of items * Number of response choices	1.000	.20
Sample Size * Number of items * Number of response choices	1.25	.50

Single-level Person and item reliability. Table 9 represents the Person and item reliability for two data sets within the RSM single-level sampling designs. As expected, the person Root Mean Square Error (RMSE) remained stable across data conditions because it is not affected by sample size but instead by person ability estimates, which were held constant. Item RMSE decreased as sample size increased.

Table 9

Pilot Study: Person and Item Reliability and Separation Indices from a Rating Scale Model Single-Level Sampling Design

File Name	Sample size (*n*)	Number of items (*i*)	Number of response choices (*rc*)	Distribution (D)	Infit Mean Square	Outfit Mean Square	Root Mean Square Error	Reliability	Separation
Rating Scale Model-Sample Size 50									
Person	50	5	4	Normal	0.99	0.99	0.82	.61	1.26
Item	50	5	4	Normal	1.00	0.99	0.23	.86	2.48
Rating Scale Model-Sample Size 200									
Person	200	5	4	Normal	0.99	0.99	0.85	.71	1.55
Item	200	5	4	Normal	1.00	0.99	0.12	.97	5.31

The ANOVA results with absolute bias as the dependent variable and sample size (*n*) is in Table 10 below. No statistically significant absolute biased estimates were found under these data conditions based on all interaction and main effects having a *p*-value > .05. The results of levels of absolute bias were computed using SPSS 23.0 (2015).

Table 10

ANOVA Results for Absolute Bias Within the Rating Scale Model

Source	*F*	*p*-value
Intercept	0.190	.740
Sample size	7.720	.220

Note, these pilot data conditions drove necessary changes in several of the data conditions examined in this dissertation. The data conditions generated for the full study are described in detail below and presented in Table 11.

Table 11

Summary of Final Data Conditions

	Sampling Design	
	Single-Level	Multilevel
Distributional characteristics	Normal Distribution	Normal Distribution
	Mixed Distribution: 50% normal, 50% Non-Normal	Mixed Distribution: 50% normal, 50% Non-Normal
	Non-Normal Distribution (skew = 3.0, kurtosis = 7.0)	Non-Normal Distribution (skew = 3.0, kurtosis = 7.0)
Cronbach and polychoric ordinal α *	.70	.70
Target between-level Intraclass Correlation Coefficients **	N/A	.20
Sample size(s):		
Individuals *(N)*	*Sample Size* = 30, 50, 300	*Sample Size* = 30, 50
Groups (*Ng)*	N/A	*Number of groups* =10, 100
Number of items: (*I*)	*Items* = 10	*Items* = 10

Table 11 (continued)

	Sampling Design	
	Single-Level	Multilevel
Number of response choices: (*rc*)	*Response choices =5*	*Response choices = 5*
For Rating Scale Model:		
Person reliability*	.7	.7
Total number of data conditions	9 = 3 (distributions) X 3 (sample sizes) X 1 (item choice) X 1 (response choice)	12 = 3(distribution) X 2 (*Sample Size*) X 2 (*Number of groups*) X 1 (item choice) X 1 (*response choice*) X 1 (between)
Total Reliability Coefficients in each Condition	X 3 reliability coefficients: Cronbach's α, polychoric ordinal α, and person reliability	X 3 level-1 reliability coefficients: Cronbach's α, polychoric ordinal α, and person reliability coefficient and 1 level-2 reliability coefficient (Spearman-Brown)
Total Number of Coefficients across Data Conditions	18	36

Note. All data generated represent a unidimensional model measured by polytomously scored items.

* indicates fixed parameters; ** to calculate Spearman-Brown Coefficient (Between Level)

Sampling Designs and Data Conditions for the Full Study

In a single-level sampling design, two reliability estimates (i.e., Cronbach's α, polychoric ordinal α, were examined within the CTT measurement framework and four reliability estimates (i.e., person reliability and separation (RSM), and item reliability and separation (RSM)) were examined within the Rasch IRT-RSM model for ordinal data. In a two-level sampling design, four reliability estimates (i.e., Cronbach's α, polychoric ordinal α, between-level Spearman-Brown and between-level Cronbach's α were examined within the MCFA framework and one reliability estimate (person reliability) was examined within the Rasch IRT-RSM for ordinal data.

Using Monte Carlo simulation techniques, data were generated to represent single-level models across 9 data conditions and two measurement frameworks (CTT and RSM), and two-level models across 12 data conditions and 2 measurement frameworks (MCFA and RSM). Main effects and interactions related to the levels of relative bias in reliability estimates were assessed and recommendations for clinicians, educators, and applied researchers are provided in Chapter V. In both the single and two-level models, multivariate normal and non-normal polytomous data were generated in R *psych package* for every data condition and the resulting levels of relative bias were examined (see Appendices C for a sample of the R code used to generate these multivariate single-level and multilevel data). Additionally, these generated item scores were saved in MS Excel and used to examine all reliability estimates calculated. The reliability in the population was fixed at .70, which George and Mallery (2003) and Serfling (2010) reported as an acceptable reliability coefficient. A full description of the single and two-level models is provided below and summarized in Table 11.

Single-Level Sampling Design

The single-level sampling design for all reliability coefficients was used as a baseline from which to compare the reliability estimates obtained in level-1 of the two-level model under all data conditions. Based on the results of Charter (1999), K. E. Green and Frantom (2002), Gadermann et al. (2012) Kahn (2014), J. B. Kline (1999, 2005), Linacre (1994), B. Muthén (1983), Nunnally (1978), Wright and Stone (1979), Yurdugul (2008), and Zumbo et al. (2007), three levels of sample size (N = 30, 50, 300) were examined. In addition, based on Bandelos and Enders (1996) and Sheng and Sheng (2012), three levels of distributional characteristics (normal, non-normal, and mixed) for all single-level models were examined. Since the effect of the number of items and the number of response choices on Cronbach's α, polychoric ordinal α, and reliability estimates using RSM have been tested extensively, and following the study designs of Lissitz and Green (1975) and Bandelos and Enders (1996), the number of items were held constant at ten (I = 10). Following the results of Bandalos and Enders (1996), Zumbo et al. (2007), and Lozano et al. (2008), the number of response choices were held constant at five (rc = 5). These 18 conditions represent a completely crossed 3 X 3 X 1 X 1 X 2 design (three distributional characteristics, 3 levels of sample size, 1 number of items, and 1 number of response choices, across 2 frameworks)

Two-Level Sampling Design

The two-level model represents a multilevel data structure where the individual is modeled at level -1 and the group is modeled at level-2. In a completely crossed design, the level-1 data sample sizes were 30 and 50. Within the two-level model two group, or cluster sizes were generated (Ng = 10, 100) for which individuals were nested, to

examine between-level reliability estimates. To date, researchers estimating parameters in a two-level model using Monte Carlo simulation techniques usually fix the number of groups to be greater than the number of subjects per group based on design effects (Bell et al., 2008; Maas & Hox, 2005; B. Muthén & Satorra, 1995; Snijders, 2005). Further, in studies focusing specifically on reliability estimates, sample sizes were similarly fixed without providing an appropriate explanation (T. A. Brown, 2015; Clark, 2008; Geldhof et al., 2014; Raudenbush, 1994; Raykov & Penev, 2010). As stated previously, and similar to the level-1 and level-2 sample sizes tested by Clark (2008), Geldhof et al. (2014), and Raudenbush et al. (2003), individual level sample sizes were 30 and 50 and group level sample sizes were 10 and 100. Conceptually, if level-1 represents students and level-2 represents classrooms, when 30 individuals in 10 classrooms are crossed, the sample size will match the 300 individuals included in the single-level model. These 48 conditions represent a completely crossed 3 X 2 X 2 X 1 X 1 X 1 X 2 design (three distributional characteristics, two levels of individual sample size, two levels of group sample size, one number of items, one number of response choices, and one between-reliability estimate, across two frameworks). A summary of data conditions examined in previous research for both single-level and multilevel models is in Tables 12 and 13 below

Table 12

Summary of Data Condition Recommendations for Single-Level Models

Citation	Data Characteristics	Included Data Characteristics
Single-Level Samples Design		
Zumbo et al. (2007)	Sample size	1000
	Number of items	14
Gadermann et al. (2012)	Number of response choices	2, 3, 4, 5, 6, 7
	Distributional characteristics	normal (skew = 0) and non-normal (skew = -2)
	Theoretical reliability	.4, .5, .8, .9
Charter (1999)	Sample size	50, 100, 200, 300, 500, 1,000
	Theoretical reliability	.5, .6, .7., .8, .9, .95
Lissitz and Green (1975)	Sample size	50
	Number of items	10
	Number of response choices	2, 3, 4, 5, 7, 9, 14
	Item Covariance	.2, .5, .8

Table 12 (continued)

Citation	Data Characteristics	Included Data Characteristics
Bandelos and Enders (1996)	Sample size	100
	Number of items	10
	Number of response choices	3, 5, 7, 9, 11
	Inter-item correlations	.25, .5., 75
	Distributional characteristics	normal (skew = 0, kurtosis = 0)
		skew = 1.75, kurtosis = 3.75
		skew = 2.25, kurtosis = 7
		platykurtic (skew = .25, kurtosis = -1)
		leptokurtic (skew = 0, kurtosis =3)
Lozano et al (2008)	Sample size	10, 100, 200, 500
	Number of items	30
	Number of response choices	2, 3, 4, 5, 6, 7, 8, 9
	Inter-item correlations	.2, .3, .4, .5, .6, .7, .8, .9

Table 12 (continued)

Citation	Data Characteristics	Included Data Characteristics
Sheng and Sheng (2012)	Sample size	30, 50, 100, 1000
	Number of items	5. 10, 30
	Number of response choices	Not provided
	Distributional characteristics	normal (skew=0, kurtosis =0)
		platykurtic (skew = 0, kurtosis = 1.35)
		non-symmetric (skew = .96, kurtosis =.13)
		leptokurtic (skew = 2.5, kurtosis =2.5)
	Theoretical reliability	.3, ,6, .8
Linacre (1994)	Sample size	30, 50, 100
	Number of items	10, 20
Wright and Stone (1979)	Number of response choices	5

Table 13

Summary of Data Condition Recommendations for Multilevel Models

Citation	Data Characteristics	Included Data Characteristics
Geldhof et al. (2014)	Level-1 sample (per group)	2, 15, 30
	Level-2 sample size (number of groups)	50, 100, 200
	Number of items	6
	Theoretical reliability	0.85
	ICC	.05, .25, .50, .75
	Factor Loadings	.6, .7, .8
Huang and Cornell (2016)	Level-1 sample (per group)	19
	Level-2 sample size (number of groups)	423
	Number of items	9

Table 13 (continued)

Citation	Data Characteristics	Included Data Characteristics
Little (2013)	Level-1 sample (per group)	18,255
	Level-2 sample size (number of groups)	2,104
	Level-3 sample size	53
	Number of items	12
	Number of response choices	5
	ICC	.20, .25
Raykov (2010)	Level-1 sample (per group)	12, 19
Raykov and Penev (2010)	Level-2 sample size (number of groups)	10, 19
T. A. Brown (2015)	Level-1 sample (per group)	10
	Level-2 sample size (number of groups)	85
	Number of items	5
	ICC	>.10

Simulation Procedures and Building the Models

Monte Carlo data simulation refers to generating samples from a specified underlying population distribution based on user provided information about the distribution and structure of a model (Bentler, 2006). In this dissertation, multivariate Monte Carlo simulation was conducted to generate 1,000 data sets for every type of reliability estimate described previously using polytomously scored items measuring one latent trait under the varying conditions outlined above and detailed below.

Within the Classical Test Theory and Multilevel Confirmatory Factor Analysis Frameworks

Cronbach's α and polychoric ordinal α are grounded in CTT where these reliability estimates are a function of the number of items on a given assessment, the average covariance between item-pairs, and the variance of the total score (Cronbach, 1951) and error is a unitary construct. I estimated Cronbach's α and polychoric ordinal α under each of the above data conditions, focusing specifically on multilevel models with non-normal data distributions, as recommended by Raykov and Penev (2010), Geldhof et al. (2014), and Huang and Cornell (2016).

Generating Single-Level Data Sets

Using Monte Carlo simulation techniques, 1,000 replications of every data condition were generated to represent item scores generated from either a multivariate normal distribution or a multivariate non-normal distribution in R using the *psy*, *psych*, *MASS*, and *sn* packages. Following the advice from van de Eijk and Rose (2015), scores on the items from the aforementioned distributions were generated to represent three distinct item distributions: (a) all items representing a normal distribution with a skew = 0

and kurtosis = 0, (b) all items representing a skewed and leptokurtic (non-normal) distribution with a skew = 1.75 and kurtosis = 3, and (c) five items drawn from the underlying normal distribution and five items drawn from the underlying non-normal distribution, representing a 50/50 "mixed distribution." Using the *mvnorm* or *dmultinom* functions, a vector of means was created based on the sample size for ten items with five response choices per item and both a Pearson correlation matrix and polychoric correlation matrix were specified with the diagonals of the correlation matrix = 1. Sample data reflecting the underlying population data conditions of sample size (*n)*, 10-items with 5-response choices each were generated from the specified population distributions with a fixed reliability estimate = .70. This provided a baseline from which to draw conclusions regarding relative bias in reliability estimates and is supported by results from Raudenbush and Bryk (2002) and Geldhof et al. (2014).

Estimating Reliability in a Single-Level Sampling Design

Cronbach's α was estimated in a single-level sampling design across all data conditions in R by specifying a Pearson correlation matrix (cor.mat) and using the *psy, psych, MASS and MBESS* packages in R (see Appendix A for an example of the R code).

A polychoric ordinal α and the standard error of polychoric ordinal α was examined across all data conditions using the multivariate normal and multinomial non-normal distributions generated previously. Following the example R code provided by Gadermann et al. (2012), a polychoric correlation matrix was created using R *Commander* and downloading the *psych* package (Revelle, 2011) as well as the *AGPArotation* package (Bernaards & Jennrich, 2005). It is important to note that R simultaneously estimates polychoric correlations from the entire data matrix rather than

using pairwise comparisons. In a single-level sampling design in R, using the polychoric correlation matrix created in R, polychoric ordinal α was estimated based on Cronbach's α and the SE_α (see Appendix B).

All 1,000 Cronbach's α coefficients and 1,000 polychoric ordinal α coefficients under every data condition were calculated in R and exported to an MS Excel file in order to (a) determine how often the known population reliability of .70 fell within the 95.0% sample confidence interval as recommended by Raykov and Penev (2010), Geldhof et al. (2014), and Wu and Zumbo (2008); and (b) following Geldhof et al. (2014) and T. A. Brown (2015), calculate relative bias in reliability estimates across all data conditions . The MS Excel file for the single-level model contains 9,000 Cronbach α and 9,000 polychoric ordinal α coefficients (1,000 iterations or data sets X 3 distributions, X 3 sample sizes).

Generating Multilevel Data Sets

Multilevel confirmatory factor analysis (MCFA) was computed using a modified version of the *multilevel* package in R. First, I generated two-level data sets with 1,000 iterations per data condition in R by using the *psych* and *lme4* packages and the *sim.multilevel* function in R and the same seed for each condition used in the single-level data generation for a multivariate normal and multivariate non-normal distributions and exported them to MS Excel files to examine their properties. Next, the MCFA model was built using a modified version of the *multilevel* package in R by specifying the thresholds, and analyzing the results within the MCFA framework (see Appendix C for an example R code used to generate the two-level data structure using the Pearson Correlation matrix and polychoric correlation matrix respectively). The within-group Cronbach and

polychoric α reliability coefficients were fixed to .70 as recommended by Hox (2002). The target between-group ICC was fixed at .20, as recommended by Ludtke et al. (2008), and the factor loadings were fixed to .6 for 5 items and .8 for five items, as presented in Geldhof et al. (2014).

Estimating Reliability in a Two-Level Sampling Design

As previously discussed, estimating Cronbach's α and polychoric ordinal α in a multilevel sampling design required an MCFA. Using MCFA, the latent trait of interest was estimated separately from item responses. In other words, MCFA "separated person traits from specific items given, and item properties from specific persons in a sample" (Templin & Bradshaw, 2013, para. 12). While MCFA is more closely related to IRT than CTT, Geldhof et al. (2014) explained that the MCFA model can be used to decompose observed item scores into "components related to each individuals' within-cluster average true score . . . as well as each individuals' true deviation from the cluster average" (p. 75). These sources of decomposed variance represent a ratio of true variance to total score variance as found in CTT. The MCFA approach used in this dissertation was represented by the within-and between-level model

The factor loading matrices ($\Lambda_{w,}$ Λ_B) and cluster level means (μ) are fixed effects while μ_B refers to the random intercepts in the Y variable (note that the thresholds used due to polytomously scored items are based on Equation 31 below). Using this approach, and fixing the first factor loading to 1 as recommended by T. A. Brown (2015), random intercepts of the Y variable are allowed to vary.

Equation 31 represents the relationship between a latent response distribution, y*, and an observed ordinal distribution, y and is applied to the level-1 (within) MCFA

model to indicate thresholds for observed polytomous data (recall thresholds = the number of response categories -1; Flora & Curran, 2004; Little, 2013):

$$y = c, \text{ if } \tau_c < y^* < \tau_{c+1} \quad (31)$$

with thresholds τ as parameters defining the categories c = 0, 1, 2, …, C -1 and where τ_0 = negative infinity and τ_c = positive infinity. Little (2013) demonstrated a multilevel factor model with polytomous response variables at level-1 and continuous random intercepts at level-2. Figure 4 is modified to represent a four-item model with one factor from Little's two-level model.

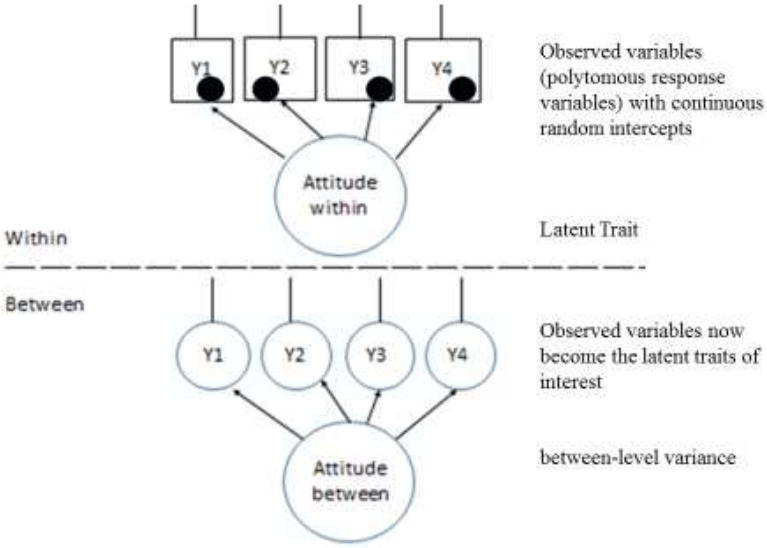

Figure 4. An example of a four-item two-level polytomous factor model. The dashed line represents the division between the between-level (level-2) model and the within-level (level-1) model. At level-1, the solid black circles at the end of the arrows for Items 1 to 4 represent observed polytomous response variables referred to as y1 to y4 in level-2. Adapted from Little (2013).

Estimating Reliability in the Two-Level Model

As with single-level data, all 1,000 level-1 (within level) Cronbach's α and polychoric ordinal α estimates were calculated in R and exported to an MS Excel file. For the between-level of the two-level model, 1,000 between level ICCs under every data condition were calculated and used to compute the Spearman-Brown Coefficient for both measurement frameworks. These calculations were then exported to MS Excel spreadsheets to (a) determine how often the known population reliability of .70 fell within the 95% confidence interval as recommended by Raykov and Penev (2010), Geldhof et al. (2014), and Wu and Zumbo (2008), (b) adopt Dedrick and Greenbaum's (2010) use of the Spearman-Brown Prophecy Formula to assess between-level reliability for both CTT and Rasch RSM frameworks, and (c) following the recommendations of Geldhof et al. (2014) and T. A. Brown (2015), calculate relative bias in all reliability estimates across all data conditions.

Within the Rasch Item Response Theory Framework

While it is possible to generate the RSM data and corresponding results in R, most social science researchers will not have the requisite programming experience to conduct their RSM analysis in R and may feel more comfortable using Winsteps. Therefore, to provide comparable results across all measurement frameworks, regardless of the software used in the analysis, the data generated previously in R for both the single and two-level models was imported into Winsteps to generate results for this dissertation. See Appendix D for R code for the single-level person reliability model. See Appendix C

for sample R code used to generate the multilevel data sets and Appendix E for Winsteps codes to estimate multilevel person reliability.

Estimating Reliability Within the Rasch Item Response Theory Framework

The single-level model. I assessed and reported both person and item reliability and separation under all nine data conditions in a single-level model, as recommended by Kamata (2001) and Raudenbush et al. (2003), using the RSM described previously. Wright and Stone (1999) and Wright and Masters (1982) illustrated that person and item reliability (which they call person and item separation reliability) is comparable to KR-20 or Cronbach's α when corrected by degrees of freedom. Item reliability and person and item separation indices were reported, but not compared in single-level models and neither reported nor compared in multilevel models. This was due to the properties of these indices. For any IRT model, individual person ability and item standard errors of measurement can be computed and assessed; however, for CTT, standard errors can only be computed and assessed for measures at the group mean of person ability, and not for individual persons or items. Therefore, only person reliability estimates in Rasch IRT models have an equivalent coefficient in which to make comparisons in CTT. Equation 32 represents person reliability (Wright & Stone, 1999):

$$\text{Person reliability} = 1-(MSEp/\sigma_p^2) \quad (32)$$

Where *MSEp* is the sample mean square person error and σ_p^2 is the sample person variance. Equation 33 represents person separation (Wright & Stone, 1999):

$$\text{Person separation} = \sqrt{\frac{1-(MSEi/\sigma i2)}{1-(1-(MSEi/\sigma i2)}} \quad (33)$$

Which represents the ratio of the person reliability estimate to 1- the person reliability estimate. Consider person reliability =.8. Person separation would then = .20 which means that only two levels of person ability can be consistently identified suggesting the person sample is too homogenous (Bond & Fox, 2014).

Equations 34 and 35 represent item reliability and separation (Wright & Stone, 1999):

$$\text{Item reliability} = 1-(MSEi/\sigma i^2) \quad (34)$$

Where *MSEi* is the sample mean square item error and σi^2 is the sample item variance.

$$\text{Item separation} = \sqrt{\frac{1-(MSEi/\sigma i2)}{1-(1-(MSEi/\sigma i2)}} \quad (35)$$

Which represents the ratio of the item reliability estimate to 1- the item reliability estimate.

The multilevel sampling design. The focus of this dissertation was to compare bias in reliability estimates across two measurement frameworks and single and multilevel models. Therefore, for the Rasch-IRT (RSM) model, only person reliability was assessed at level-1 and Spearman-Brown Prophecy coefficients were assessed at level-2 across all data conditions of 3 reliability level-1 coefficients X 1 measurement framework (Rasch-IRT: RSM) X 2 level-1 sample sizes (N =30 and N = 50) X 2 level-2 sample sizes (n = 10 and n = 100), and X 3 distributions (normal, mixed, and non-

normal) for a total of 36 data conditions in a multilevel model, as recommended by Kamata (2001) and Raudenbush et al. (2003), in RSM. As with the single-level RSM model, the person ability and item difficulty parameters were fixed with a mean = 0 and *SD* = 1, as stated previously. The between level ICC target was > .20 to match the two-level model specifications used in MCFA. Once these data-sets were generated under every data condition they were saved to MS Excel spreadsheets and imported into Winsteps for analysis. Level-2 comparisons were made by calculating between-level variance within the CTT and Rasch (RSM) frameworks and assessed by conducting a factorial ANOVA and computing and comparing Spearman-Brown Coefficients across data conditions.

Final Data Conditions Examined

In the single-level sampling design, 54 reliability coefficients (i.e., 9 data conditions X 6 types of reliability) X 1,000 replications per cell totaling 54,000 reliability coefficients, were generated along with 54,000 standard errors of reliability, 95% confidence intervals (Lower Level and Upper Level), and relative biases for a total of 270,000 cells. In the multilevel sampling design, 60 reliability coefficients (i.e., 12 data conditions X 5 types of reliability) X 1,000 replications per cell, totaling 60,000 reliability estimates, were generated, along with 60,000 standard errors of reliability, 95% confidence intervals (Lower Level and Upper Level), and relative biases for a total of 300,000 cells. All data were saved in MS Excel for further analysis and the results for these 570,000 cells are presented in Chapter IV.

Summary of Final Data Conditions

The final data conditions outlined below are consistent with conditions selected by previous methodological researchers using Monte Carlo simulation techniques to address issues related to parameter estimation, specifically reliability estimates. In addition, these data conditions represent more realistic sample sizes and hold the number of items and number of response choices constant to elucidate necessary data conditions for relative bias found in reliability estimates. Table 11, presented previously, shows all data conditions for both single-level and multilevel sampling designs. Under each sampling design, all data conditions were crossed, and main effects and interactions were assessed. Since the estimation of relative bias of reliability coefficients in a multilevel model with non-normal data was recommended, but not published, decisions regarding the distribution of data were based on Bandelos and Enders' (1996) Monte Carlo simulation study assessing the effects of non-normal data on the number of response categories in a single-level sampling design and the recommendations of T. A. Brown (2015), Geldhof et al. (2014), and Huang and Cornell (2016) regarding the degree of non-normal data in multilevel models.

Following previous results and recommendations, I selected one normal distribution of all items (skew = 0, kurtosis = 0), one non-normal leptokurtic distribution of all items (skew = 1.75, kurtosis = 3.0), and, based on the methods presented by van de Eijk and Rose (2015), I added what I refer to as a "mixed distribution" of items (50% of items were drawn from a normal distribution and 50% of items were drawn from the skewed and leptokurtic distribution as described previously). As with my pilot study, using the *mvnorm* or *dmultinom* functions, a vector of means was created based on the

sample size for ten items with five response choices per item and both a Pearson correlation matrix and polychoric correlation matrix were specified with the diagonals of the correlation matrix = 1. Sample data reflecting the underlying population data conditions of sample size (*n)*, 10-items with 5-response choices each were generated from the specified population distributions with a fixed reliability estimate = .70, and skew and kurtosis were modified (skew = +/- 3.0, kurtosis = +/- 7.0) for the non-normal and mixed data distributions to better reflect the thresholds at which non-normality increases bias in reliability coefficients. These types of data could be seen if the first five items are 'easy' for respondents to endorse while the next five items are more 'difficult' for respondents to endorse, causing a range of normally distributed scores combined with a cluster of low scores.

Finally, Zhang (2010) presented a simulation study for the Rasch family of models where ability and difficulty parameters were fixed to a mean = 0, *SD* = 1, and person and item reliability were estimated. Zhang did not estimate person or item separation. Based on the fixed person reliability = .70 and the person ability variance = 1, the person separation index is expected to be 1.52. Based on the fixed item difficulty variance = 1, the item reliability estimate is expected to be ≥ .95, and the separation index is expected to be 4.35 in RSM. The reliability separation estimate is lower than the person separation = 2 and item separation = 7 recommended by Bond and Fox (2014), Kim and Feldt (2010), Linacre (2004, 2014), and Rutkowski & Svetina (2013) but represent more realistic estimates based on the data conditions affecting relative bias in reliability estimates. For example, Linacre (2014) explained that person reliability and separation estimates depend on person ability variance, number of items, and number of

response categories, with higher levels of each resulting in higher person reliability and separation. Item reliability and separation depend on item difficulty variance and person sample size with higher levels of each resulting in higher item reliability and separation. Person separation and person and item reliability are presented in the single-level sampling design results but, as explained previously, not used when comparing bias between CTT and Rasch-RSM multilevel models.

Data Analysis

To answer the four research questions posed in this dissertation, the data analysis consisted of first computing relative bias in reliability coefficients across all data conditions, sampling designs, and measurement frameworks and then analyzing bias as outlined below.

Data representing 18 reliability coefficients in the single-level model and 36 reliability coefficients in the multilevel model were generated in R, exported to MS Excel, imported into Winsteps for RSM models and loaded into a modified R multilevel package for MCFA models. Once all reliability coefficients were computed and exported to MS Excel, relative bias was calculated using Equation 36 shown below where "known reliability" were the Spearman-Brown Coefficients computed from the ICC's:

$$Relative\ Bias = \frac{known\ reliability - sample\ reliability}{known\ reliability} \tag{36}$$

Further analysis regarding relative bias included (a) determining the proportion of known reliability coefficients falling within the sample confidence intervals with ≥ 95% deemed acceptable; (b) assessing whether reliability estimates were over-or-underestimated based on the based on research conducted by B. Muthén and Kaplan, (1985) and Geldhof et al.

(2014), where relative bias ≤ 10% was deemed acceptable; and (c) examining the behavior of the standard errors, which were expected to decrease as sample size increased, regardless of the distribution of data or sampling design.

Finally, factorial ANOVAs across data conditions were conducted in SPSS 24.0 with the dependent variables of reliability coefficients, standard errors, and percentage of relative bias and the independent variables of the type of reliability coefficient, sample size, and distributional characteristics indicated as the fixed factors with varying levels. All two-way interaction effects between factors were examined for significance first, and if significant, simple main effects were examined, along with the interaction graphs. Main effects were then examined by splitting the factors into groups and identifying the separate levels of each group based on sampling design. For statistically significant results, eta squared (η^2) was calculated and examined for every simple main effect and main effect. Partial eta squared was not used as it tends to overestimate the true effect in complex models. The criteria for a medium effect size was any $\eta^2 > .06$ as recommended by Cohen (1977, 1988) and by Thompson (2007) who conducted a simulation study of effect sizes. Chi-square tests were conducted to assess the percentage of bias either overestimated or underestimated across all data conditions. As presented by T. A. Brown (2015), Geldhof et al. (2014), and Huang and Cornell (2016), any *p*-value ≤ .05 indicated a statistically significant level of bias in reliability estimates. Results are presented in Chapter IV.

Chapter Summary

The research questions were formulated based on the specific gaps in the literature pertaining to reliability estimation in a two-level model and recommendations

for future research posed by Geldhof et al. (2014), Zumbo et al. (2007), and Gadermann et al. (2012). The decisions regarding the data conditions I held constant (unidimensionality of the latent trait, polytomously scored items, number of items, number of response choices, reliability coefficients, ICCs and person and item parameters) were informed by research conducted by Bandelos and Enders (1996), Bond and Fox (2014), Gadermann et al. (2012), Lozano et al. (2008), Little (2013), Maas and Hox (2005), B. Muthén and Muthén (2000), Sheng and Sheng (2012), and Zumbo et al. (2007) as well as recommendations made by Linacre (2014), Nunnally and Bernstein (1994), Yurdugul (2008), and my dissertation committee. The distribution of data was inspired by Bandelos and Enders (1996) and Sheng and Sheng (2012) as well as advice. Finally, sampling designs were modeled after Maas and Hox (2005).

The methods employed to answer the four research questions were supported by T. A. Brown (2015), Geldhof et al. (2014), Huang and Cornell (2016), van de Eijk and Rose (2015), and others. A simulation study was chosen to provide control over the data characteristics and allow for comparisons of relative bias in reliability estimates within- and- across measurement frameworks. Guidelines for estimating and measuring relative bias in reliability coefficients were based on Geldhof et al. (2014) and Huang and Cornell (2016).

CHAPTER IV

RESULTS

Using reliability estimates computed for each data condition, sampling design, and measurement framework outlined in Chapter III, I examined the proportion of relative bias and associated confidence intervals and reported the results as percentages. Recall that relative bias is a measure of the proportion of bias as it relates to the known reliability coefficient in each data condition. One thousand replications for every data condition, across single-level and multilevel models, and two measurement frameworks, were simulated and exported into MS Excel spreadsheets. Reliability coefficients and standard errors of measurement within each data condition, sampling design, and measurement framework were then averaged and 95% confidence intervals and bias were computed and reported.

Presentation of Results

Results are presented by research question and hypothesis. Tables 14 and 15 outline how the results of this dissertation are organized, and Table 16 illustrates the comparisons across measurement frameworks, data conditions, and data structures.

To untangle the results, first I adopted Charter's (1999) and Sijtsma's (2009) recommendations to assess and report, not only the actual reliability estimates, but the standard errors and corresponding confidence intervals, and presented the results in Table 17. Based on data conditions and type of reliability estimate, the proportions of coefficients falling outside the 95% reliability confidence intervals are presented in Table

18. Second, I followed the guidelines outlined by Muthén (1987) and Geldhof et al. (2014) who stipulated that relative bias ≤ 10% is acceptable, and present these results in Table 19. Finally, I conducted tests of hypotheses for all three average reliability coefficients and their corresponding standard errors and percentages of bias across all data conditions in IBM SPSS Statistics for Windows, Version 24.0 and present the results in Table 20.

Recall that item reliability and person and item separation indices are reported in Winsteps and R, but neither reported nor compared within or across measurement frameworks in this dissertation. This decision was made based on two important factors: (a) Comparing these indices within the Rasch framework is beyond the scope of this dissertation, which focused only on measures of reliability analogous to Cronbach's α, and is recommended for future research and (b) the properties of these indices do not permit direct comparison. For any Rasch model, cluster person reliability and item standard errors of measurement can be computed and assessed; however, for CTT, standard errors can only be computed and assessed for measures at the cluster mean of person ability, and not for cluster persons or items. Therefore, only person reliability estimates across frameworks were compared for both single and level-1of multilevel models. Level-2 comparisons were made by calculating between-level variance within the CTT and Rasch (RSM) frameworks and assessed by using intraclass correlation coefficients (ICC's) to calculate Spearman-Brown coefficients across data conditions.

Table 14

Presentation of Single-Level Results

Single-Level Bias by Measurement Framework and Data Condition		
Measurement Framework	Distribution	Reliability
Classical Test Theory	Normal	Cronbach's α
		Polychoric ordinal α
	Mixed	Cronbach's α
		Polychoric ordinal α
	Non-Normal	Cronbach's α
		Polychoric ordinal α
Item Response Theory-Rating Scale Model	Normal	Person Reliability
		Person Separation
	Mixed	Person Reliability
		Person Separation
	Non-Normal	Person Reliability
		Person Separation

Table 15

Presentation of Multilevel Results

Multilevel Bias by Measurement Framework and Data Condition		
Measurement Framework	Distribution	Reliability
Classical Test Theory	Normal Distribution	Cronbach's α (Within groups)
		Intraclass Correlation and Spearman-Brown Coefficient (Between Groups)
		Polychoric ordinal α (Within Groups)
		Intraclass Correlation and Spearman-Brown Coefficient (Between Groups)
	Mixed	Cronbach's α(Within groups)
		Intraclass Correlation and Spearman-Brown Coefficient (Between Groups)
		Polychoric ordinal α (Within Groups)
		Intraclass Correlation and Spearman-Brown Coefficient (Between Groups)
	Non-Normal Distribution	Cronbach's α (Within groups)
		Intraclass Correlation and Spearman-Brown Coefficient (Between Groups)
		Polychoric Ordinal α (Within Groups)
		Intraclass Correlation and Spearman-Brown Coefficient (Between Groups)

Table 15 (continued)

Multilevel Bias by Measurement Framework and Data Conditions		
Measurement Framework	Distribution	Reliability
Item Response Theory (Rating Scale Model)	Normal	Person Reliability
		Intraclass Correlation and Spearman-Brown Coefficient (Between Groups)
	Mixed	Person Reliability
		Intraclass Correlation and Spearman-Brown Coefficient (Between Groups)
	Non-Normal	Person Reliability
		Intraclass Correlation and Spearman-Brown Coefficient (Between Groups)

Table 16

Comparison of Bias Across Measurement Frameworks, Data Conditions, and Data Structures

Comparison of Sampling Designs and Measurement Frameworks	Reliability Coefficients Compared
Classical Test Theory Single and Multilevel Models	Comparing Cronbach's α and polychoric ordinal α in a single-level Classical Test Theory model to level-1 of a multilevel model
Classical Test Theory Single-Level and Item Response Theory (Rating Scale Model) Single-Level Models	Comparing Cronbach's α and polychoric ordinal α in single-level Classical Test Theory model to person reliability in an Item Response Theory (Rating Scale Model) single-level model
Classical Test Theory Multilevel and Item Response Theory (Rating Scale Model) Multilevel Models	1. Comparing Cronbach's a in level-1 of Classical Test Theory framework and person reliability in level-1 of an Item Response Theory (Rating Scale Model) framework. 2. Calculating and Comparing the Spearman-Brown Coefficients in level-2 of the Classical Test Theory and Item Response Theory (Rating Scale Model) frameworks

Table 17

Single-Level Sample Reliability Coefficient Results

Sample Size	Average Sample Reliability*	Average Sample Reliability SE	95% Confidence Interval	
			Average Lower Level	Average Upper Level
Cronbach's a:				
All Data Normally Distributed				
30	.682	.091	.500	.864
50	.690	.070	.551	.830
300	.703	.025	.655	.753
Mixed Data Distribution				
30	.809	.056	.696	.921
50	.857	.035	.787	.930
300	.820	.144	.533	1.107
Non-Normal Distribution				
30	.608	.153	.302	.914
50	.607	.100	.408	.806
300	.579	.033	.513	.645
Polychoric Ordinal a				
All Data Normally Distributed				
30	.662	.100	.463	.862
50	.669	.086	.497	.841
300	.689	.031	.627	.751

Table 17 (continued)

Sample Size	Average Sample Reliability	Average Sample Reliability SE	95% Confidence Interval	
			Average Lower Level	Average Upper Level
Mixed Data Distribution				
30	.809	.059	.692	.926
50	.783	44.000	.647	.918
300	.820	.016	.787	.853
Non-Normal Distribution				
30	.367	.211	-.054	.789
50	.764	.081	.601	.927
300	.622	.083T	.456	.788
Person Reliability (Rating Scale Model)				
All Data Normally Distributed				
30	.797	.060	.677	.917
50	.846	.012	.822	.869
300	.848	.012	.825	.871
Mixed Data Distribution				
30	.590	.310	-.030	1.21
50	.680	.380	-.080	1.44
300	.610	.460	-.310	1.530

Table 17 (continued)

Sample Size	Average Sample Reliability	Average Sample Reliability SE	95% Confidence Interval	
			Average Lower Level	Average Upper Level
Non-Normal Distribution				
30	.430	.940	-1.450	2.310
50	.410	.830	-1.250	2.070
300	.380	.850	-1.310	2.080

* Population reliability coefficients fixed to .70.

Table 18

Absolute Value and Percentage of Bias Across Type of Reliability, Data Distribution, and Sample Size

Sample Size	Average Reliability Coefficient Relative Bias	Percentage of Bias ≥ 10%	Average Bias ≥ 10%: Percentage Underestimated	If Average Bias ≥ 10%, Percentage Overestimated
Cronbach's α:				
	All Data Normally Distributed			
30	.104	39.50%	31.60%	7.90%
50	.076	26.90%	26.90%	0.00%
300	.029	0.50%	0.50%	0.00%
	Mixed Data Distribution			
30	.159	78.20%	0.50%	77.70%
50	.172	99.80%	0.00%	99.80%
300	.225	98.50%	0.70%	97.80%
	Non-Normal Distribution			
30	.175	55.90%	53.30%	2.60%
50	.149	56.00%	51.70%	4.30%
300	.173	94.40%	94.30%	0.10%
Polychoric ordinal α				
	All Data Normally Distributed			
30	.113	41.20%	36.10%	5.10%
50	.082	29.90%	29.90%	0.00%
300	.032	2.00%	2.00%	0.00%

Table 18 (continued)

Sample Size	Average Reliability Coefficient Relative Bias	Percentage of Bias ≥ 10%	Average Bias ≥ 10%: Percentage Underestimated	If Average Bias ≥ 10%, Percentage Overestimated
		Mixed Data Distribution		
30	.162	79.00%	1.70%	77.30%
50	.172	68.60%	0.00%	68.60%
300	.136	99.70%	4.50%	95.20%
		Non-Normal Distribution		
30	.475	96.60%	96.50%	0.10%
50	.129	62.40%	10.40%	52.00%
300	.111	55.90%	55.80%	0.10%
Person Reliability (Rating Scale Model)				
		All Data Normally Distributed		
30	.163	84.00%	0.00%	84.00%
50	.208	100.00%	0.00%	100.00%
300	.212	100.00%	0.00%	100.00%
		Mixed Data Distribution		
30	.157	18.00%	0.00%	18.00%
50	.029	2.00%	0.00%	2.00%
300	.000	0.00%	0.00%	0.00%

Table 18 (continued)

Sample Size	Average Reliability Coefficient Relative Bias	Percentage of Bias ≥ 10%	Average Bias ≥ 10%: Percentage Underestimated	If Average Bias ≥ 10%, Percentage Overestimated
Non-Normal Distribution				
30	.386	100.00%	100.00%	0.00%
50	.414	100.00%	100.00%	0.00%
300	.457	100.00%	100.00%	0.00%

Table 19

Results of Reliability Estimates and Standard Errors for Tests of Hypotheses (ANOVA's 1 and 2)

DV	IV	Levels	*F* statistic	*p*-value	Post-hoc/Additional Analysis Results	Effect Size ($\eta 2^{**}$)
Reliability Coefficients	Type of Reliability	Cronbach's α, polychoric ordinal α, person reliability	.630	.541		
	Sample Size	30, 50, 300	.413	.666		
	Distribution	normal, mixed, non-normal	13.42	< .0001*	Bonferonni: normal to non-normal (p =.001) and mixed to non-normal (p < .0001)	.582
	Measurement Framework	Classical Test Theory or Item Response Theory (Rating Scale Model)	1.232	.278		
Standard Errors of Measurement	Type of Reliability	Cronbach's α, polychoric ordinal α, person reliability	7.85	.002*	Bonferonni: Cronbach α to person reliability (p=.007) and polychoric ordinal α to person reliability (p=.006)	.395
	Sample Size	30, 50, 300	.056	.946		
	Distribution	normal, mixed, non-normal	3.482	.047*	Bonferonni: normal to non-normal (p = .045)	.225
	Measurement Framework	Classical Test Theory and Item Response Theory (Rating Scale Model)	4.044	< .0001*		.837

* denotes a significant result at the a =.05 level of significance; ** eta squared is only reported when statistically significant differences exist

Table 20

Results of Relative Bias and Percentage/Direction of Relative Bias for Tests of Hypotheses (ANOVA's 3 and 4)

DV	IV	Levels	*F* statistic	*p*-value	Post-hoc/Additional Analysis Results	Effect size (*η2***)
Absolute Value of Relative Bias in Reliability Estimates	Type of Reliability	Cronbach's α, polychoric ordinal α, person reliability	625.865	< .0001*	Bonferonni: Cronbach's α, to polychoric ordinal α, Cronbach's α to person reliability, polychoric ordinal α to person reliability (all $p < .0001$)	.394
	Sample Size	30, 50, 300	.167	.683		
	Distribution	normal, mixed, non-normal	181.579	< .0001*	Bonferonni: normal to mixed, normal to non-normal, mixed to non-normal (all $p < .0001$)	.339
	Measurement Framework	Classical Test Theory and Item Response Theory (Rating Scale Model)	75	< .0001		.239

Table 20 (continued)

DV	IV	Levels	*F* statistic	*p*-value	Post-hoc/Additional Analysis Results	Effect size (*η2***)
Proportion of Bias Underestimated	Type of Reliability	Cronbach's α, polychoric ordinal α, person reliability	.008	.992		
	Sample Size	30, 50, 300	.070	.933		
	Distribution	normal, mixed, non-normal	23.697	< .0001*	Bonferonni: normal to non-normal (p = .001) and mixed to non-normal (p < .0001)	.736
	Measurement Framework	Classical Test Theory and Item Response Theory (Rating Scale Model)	.065	.949		

Table 20 (continued)

DV	IV	Levels	*F* statistic	*p*-value	Post-hoc/Additional Analysis Results	Effect size (*η2***)
Proportion of Bias Overestimated	Type of Reliability	Cronbach's α, polychoric ordinal α, person reliability	.031	.970		
	Sample Size	30, 50, 300	0.273	.764		
	Distribution	normal, mixed, non-normal	23.697	< .0001*	Bonferonni: normal to non-normal (p = .002) and mixed to non-normal (p < .0001)	.753
	Measurement Framework	Classical Test Theory and Item Response Theory (Rating Scale Model)	.212	.836		

* denotes a significant result at the a =.05 level of significance; ** eta squared is only reported when statistically significant differences exist

A Recap of Data Conditions and Reliability Terminology

For this dissertation, using Monte Carlo simulation techniques, sample sizes and distributional characteristics were varied and levels of bias in reliability estimates were compared across single-level and two-level data structures. Standard errors of measurement for all reliability estimates as well as bias in estimates from the following reliability coefficients are reported and/or compared: Cronbach's α (CA), polychoric ordinal α (PA), person reliability (PR), and Spearman-Brown Prophecy coefficients (SB). Detailed specifications for these varying data conditions and fixed parameters are found in Chapter III of this dissertation.

Results by Research Questions and Hypotheses

Q1 In a single-level model, to what degree do data conditions (sample size and distribution of data) affect levels of bias in reliability estimates (a comparison of Cronbach's α, polychoric ordinal α, and person reliability?

H1 Bias in reliability estimates will increase under the conditions of smaller sample sizes and non-normal or mixed distributions and polychoric ordinal α and person reliability will be less biased than Cronbach's α.

Classical Test Theory Single-Level vs. Rasch Rating Scale Model Single-Level Results

Reliability and standard errors. Table 17, presented previously, represents the descriptive statistics for three average sample reliability estimates computed from 1,000 simulated data sets for each reliability coefficient in a single-level sampling design by type of reliability estimate, sample size, data distribution, and measurement framework. Note in the simulated data, Cronbach's α, polychoric ordinal α, and person reliability were fixed at .70.

A comparison of Cronbach's α and polychoric ordinal α versus person reliability in a single-level model provided four key findings. First, Cronbach's α and polychoric ordinal α provide similar results in reliability estimates and standard errors across sample sizes and normal and mixed data distributions, except under normally distributed data where polychoric ordinal α tended to underestimate reliability. Second, a review of the patterns of estimation shows that Cronbach's α and polychoric ordinal α seldom fell outside of the 95% confidence intervals under normal and mixed distributions, and often fell outside of the 95% confidence intervals under non-normal data distributions. Third, person reliability overestimated reliability under conditions of normally distributed data across sample sizes, while standard errors were comparably low to the standard errors seen in Cronbach's α and polychoric ordinal α, and underestimated reliability under conditions of mixed and non-normal distributions across sample sizes, where standard errors were found to be quite high compared to Cronbach's α and polychoric ordinal α. Fourth, an unusual pattern emerged across data distributions for polychoric ordinal α and person reliability when $N = 50$. This pattern is explored further I Chapter V.

The overestimated person reliability results were unexpected since Linacre (2012, 2014) posited that person reliability, although analogous to Cronbach's α, would tend to be slightly underestimated in the Rasch model at lower sample sizes. I found the opposite to be true in the data simulated for this dissertation, causing me to re-examine my data generation methods in both R and Winsteps. First, I reviewed how the Classical Test Theory data were generated across all data distributions and $N = 50$ to account for the anomalies in Cronbach's α and polychoric ordinal α and then how Rasch RSM data were generated across sample sizes and data distributions to look for typos or incorrect coding.

Finding no typos or incorrect coding, I considered that person reliability depended upon (a) sample ability variance, (b) the number of response choices, (c) the number of items, and (d) was independent of sample size. For the data generated for this dissertation, sample ability variance was fixed to 1.0, as recommended by Zhang (2010), the number of response choices was fixed to 5, and the number of items was fixed to 10. Second, in a standard Rasch model, items are dichotomously scored, whereas in this dissertation, items were polytomously scored based on the Rasch rating scale model, and Linacre (2017) stated that a higher number of response categories would translate into higher person reliability estimates for smaller sample sizes in normally distributed data. My results showed clear evidence that under conditions of normally distributed data, on average, person reliability, though fixed at .70, was overestimated and under mixed and non-normal data distributions was severely underestimated. Finally, the lowest person reliability estimates and highest standard errors are seen under conditions of mixed and non-normal distributions, which are lower than estimates from Cronbach's α and polychoric ordinal α. Each of these results is discussed further in Chapter V. Figure 5 is the visual representation of the three reliability coefficients (Cronbach's α [CA], polychoric ordinal α [PA], and person reliability [PR]) found in Table 17 and shows that single-level reliability coefficients are similar across data conditions for CA and PA but vary considerably under the condition of non-normality.

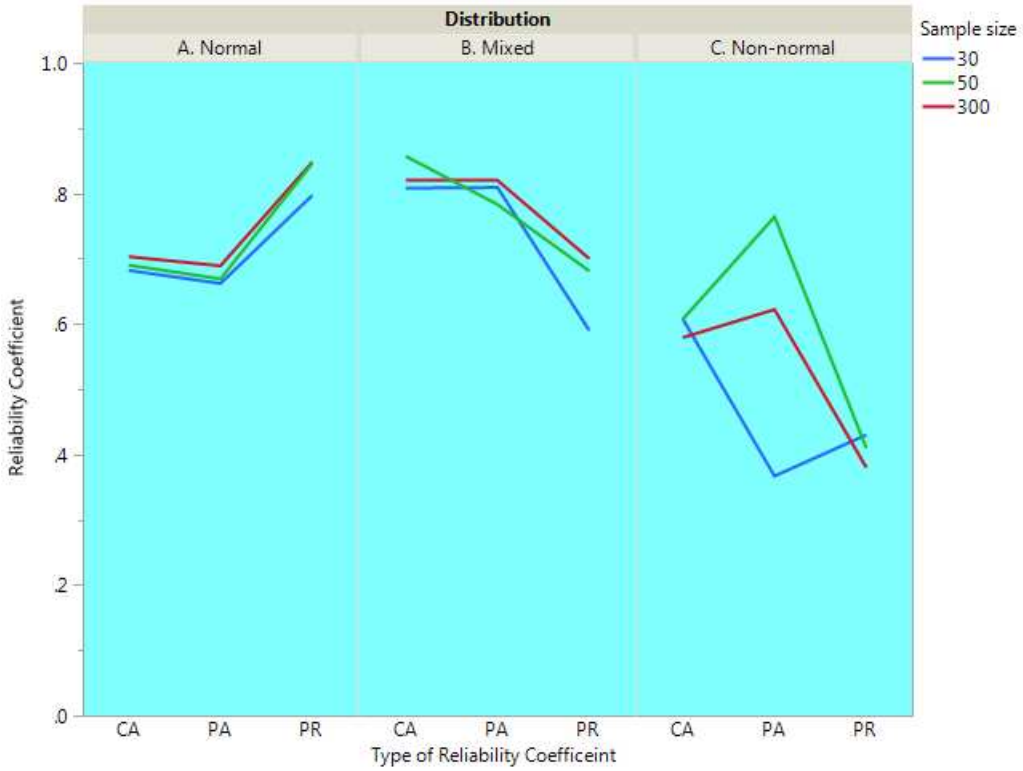

Figure 5. Single-level reliability coefficients across distributions and sample size.

Figure 6 is the visual representation of the three reliability estimate's standard errors of measurement from Table 17, which was presented previously. This figure shows that standard errors are low across Cronbach's α and polychoric ordinal α, but begin to rise for person reliability under conditions of a mixed data distribution, and rise even more so when data were non-normal. Finally, this figure shows that standard errors, while low, vary more for Cronbach's α and polychoric ordinal α than for person reliability.

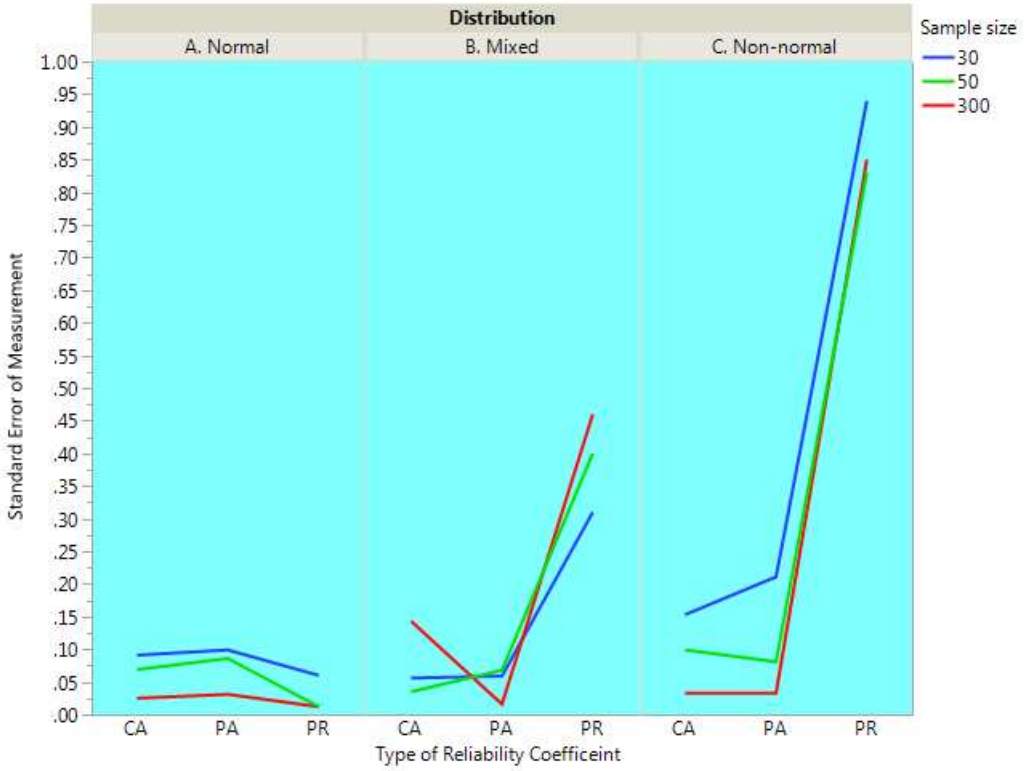

Figure 6. Single-level standard errors of reliability across distributions and sample size.

Relative bias. Table 18 shows the average absolute values of relative bias by type of reliability, data distribution, and sample size. Additionally, if the percentage of bias is ≥ 10%, then the percentage of reliability estimates either underestimated or overestimated are given.

B. Muthén (1987) and Geldhof et al. (2014) stated that relative bias < 10% was acceptable for reliability coefficients. Therefore, when the absolute value of relative bias calculated from the data generated in this dissertation was ≥ 10%, the direction (whether underestimated or overestimated) and percentage of bias was also reported.

Based on assessing the absolute value of relative bias reported in Table 18, the primary findings are that Cronbach's α and polychoric ordinal α were less biased than person reliability coefficients under normal data conditions and across sample sizes and relative bias was similar for Cronbach's α and polychoric ordinal α, across all data distributions and sample sizes, except for polychoric ordinal α under the conditions of non-normality and $N = 30$, where bias was quite high. Table 18 shows that under the condition of normally distributed data, Cronbach's α and polychoric ordinal α had a tendency to underestimate reliability in $N = 30$. This result makes sense since, according to Cronbach (1951), α represents the lower bound of reliability. Under larger sample sizes ($N = 50$ and $N = 300$), neither Cronbach's α nor polychoric ordinal α showed an average bias ≥ 10%. Comparatively, under the same normal distribution and across sample sizes, person reliability overestimated reliability 84% of the time. This result was unexpected, as previously mentioned regarding high reliability coefficients, and is explored in Chapter V.

Figure 7 shows that the average percentage of absolute relative bias based on data distribution and sample size is quite low for person reliability in a mixed distribution and quite high under smaller sample sizes and non-normal distributions across types of reliability. Overall the highest relative bias appeared to be for polychoric ordinal α under non-normal data when $N = 30$.

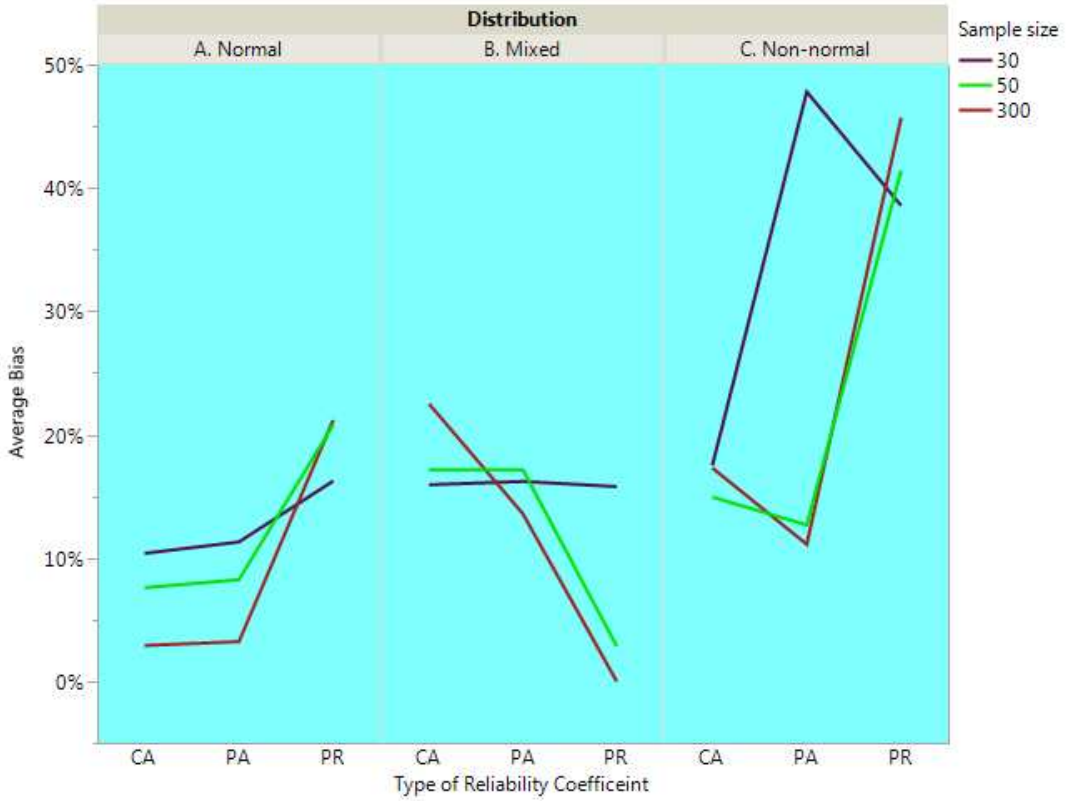

Figure 7. A visual representation of relative bias across data distributions and sample size.

Figure 8 shows the direction and percentage of absolute relative bias by distribution and sample size only if the bias recorded was ≥ 10%. For example, under the conditions of normality, Cronbach's α and polychoric ordinal α are underestimated, on average, between 30% and 36% of the time when *N*= 30, and never underestimated when *N* = 50 or *N* = 300 while person reliability is never underestimated, regardless of sample size.

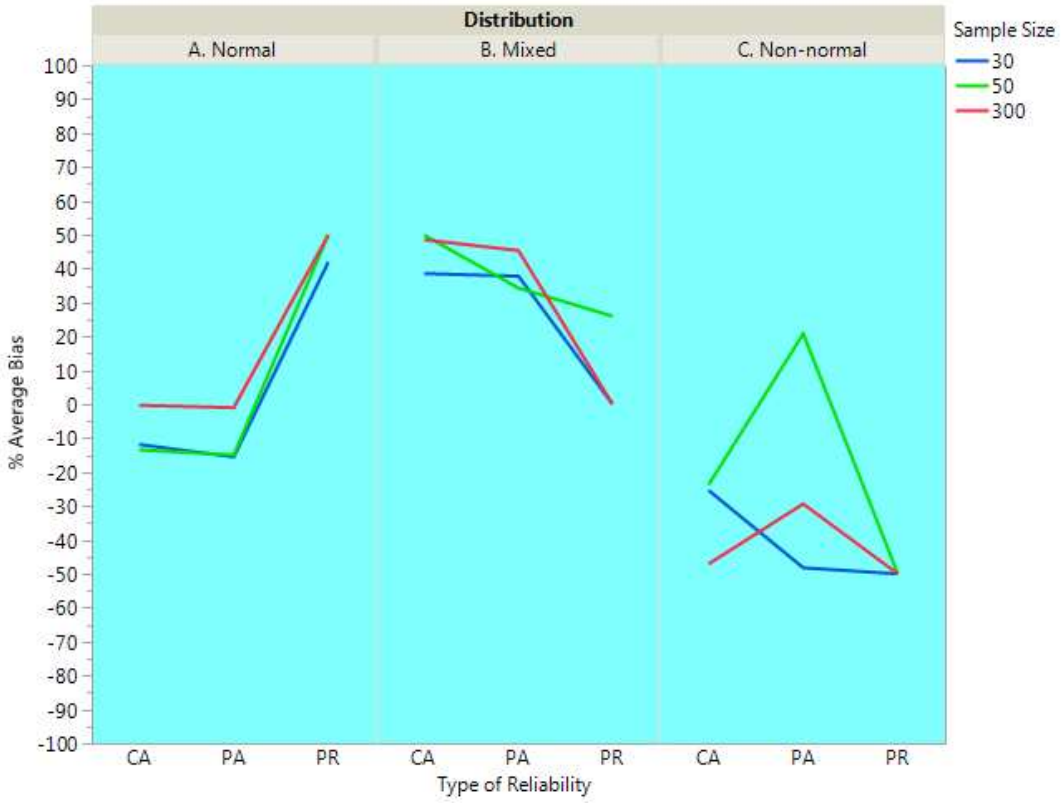

Figure 8. A visual representation of the direction and percentage of average relative bias in a single-level model across types of reliability, data distributions, and sample size.

Tests of Hypotheses for Research Question 1

Three factorial ANOVAs were conducted for the dependent variables of *reliability coefficients, standard errors of reliability coefficients,* and the *percentage of the absolute value of relative bias ≥ 10%.* One Chi-square test was conducted for the dependent variable *direction of bias (*either underestimated or overestimated), with three reliability coefficients (Cronbach's α, polychoric ordinal α, and person reliability), across three sample sizes (*N* = 30, 50, 300), three data distributions (normal, mixed, and non-normal), and two measurement frameworks (Classical Test Theory and RSM). No

statistically significant interactions were found; therefore, Tables 19 and 20 show the results by main effect. You also need to refer to Table 21 in the text prior to the table.

Table 21

Single-Level Direction of Relative Bias ≥ 10% Across Distribution and Sample Sizes (Chi-square Results)

Pearson Chi-square	*df*	Chi-square*
Distribution	2	8185.121
Type of Reliability	2	3203.19
Sample Size	1	233.649

* $p < .0001$

Table 21 shows the results of the Chi-square test assessing relative bias ≥ 10% across data distributions, types of reliability, and sample size. The data provide evidence to support my hypothesis that in single-level models, reliability estimate bias increases under the conditions of smaller sample sizes and non-normal or mixed distributions. Cronbach's α and polychoric ordinal α were the least biased under conditions of $N = 50$ and $N = 300$ when all data were normally distributed. Person reliability was the least biased when $N = 50$ and $N = 300$ when the data distribution was mixed. Polychoric ordinal α showed less bias than Cronbach's α when $N = 300$, and both Cronbach's α and polychoric ordinal α showed less bias than person reliability regardless of the data characteristics.

Data do not provide evidence that polychoric ordinal α and person reliability were less biased than Cronbach's α. Reliability coefficients computed for polychoric ordinal α differed only slightly from those computed from Cronbach's α, and most notably, person

reliability showed higher bias than Cronbach's α across data conditions. All three types of reliability coefficients showed the most bias under non-normal data distributions, with person reliability showing the most amount of bias and Cronbach's α showing the least amount of bias, though similar to polychoric ordinal α results.

Most noteworthy from the test of hypotheses was that the type of data distribution (normal, mixed, and non-normal) consistently showed statistically significant results regardless of the dependent variable (reliability coefficient, standard errors, bias, and direction of bias). For example, Cronbach's α and polychoric ordinal α showed a tendency to underestimate reliability coefficients under normal and mixed distributions and person reliability had a tendency to overestimate reliability in normal and mixed distributions. Furthermore, two additional independent variables significantly affecting the standard errors of measurement were the type of reliability coefficient and measurement framework (which by definition overlap). For all statistically significant results, η^2 is reported to provide a better understanding of the magnitude of these differences. Cohen (1977, 1988), explained that η^2 is the proportion of the *total* variability in the dependent variable that is accounted for by variation in the independent variable, analogous to R^2. He introduced a rule of thumb for interpreting η^2 for ANOVAs, where .01 represent a small effect, .06, a medium effect, and .14 a large effect. All η^2 's calculated for Q1 show large effect sizes. I interpreted these in the following way: (a) data distributions had a large effect on reliability coefficients, regardless of sample size, type of coefficient and measurement framework, (b) the type of reliability and data distributions had a large effect on standard errors of measurement, regardless of sample size, (c) the type of reliability, data distributions, and measurement frameworks had a

large effect on the percentage of the absolute value of relative bias, (d) the data distributions had a large effect on whether bias found to be ≥ 10% was underestimated or overestimated. The implications of these results are discussed in Chapter V.

Multilevel Model Bias Across Data Conditions

Q2 In a multilevel model, to what degree do data conditions (sample size and distribution of data) affect levels of bias in reliability estimates (a comparison of Cronbach's α, polychoric ordinal α, and person reliability in level-1 (within clusters) and Spearman-Brown coefficients in level-2 (between clusters)?

H2 In multilevel models, bias in reliability estimates in level-1 will increase under the conditions of smaller sample sizes and non-normal or mixed distributions and polychoric ordinal αwill be less biased than Cronbach's αand person reliability. Additionally, Intraclass correlations and Spearman-Brown coefficients will be underestimated under the conditions of smaller sample size and non-normal or mixed distributions

Recall data conditions for multilevel models included cross-sections of data conditions. For level-1 there were three distributions (normal, mixed, and non-normal), two sample sizes (N = 30 and N = 50), three within reliability coefficients across two measurement frameworks, which overlap with the three types of reliability coefficients (e.g., Cronbach's α [CTT], polychoric ordinal α [CTT], and person reliability [Rasch RSM]). For level-2 of the multilevel model there were three distributions (normal, mixed, and non-normal), two cluster sample sizes (n = 10 and n = 100) corresponding to two level-1 sample sizes, one between level coefficient (Spearman-Brown) across two three level-1 reliability coefficients and two measurement frameworks (CTT and Rasch RSM). Data for each condition were simulated with 1,000 replications for each set of conditions and average coefficients and standard errors were calculated for each set. A total of 36

level-1 reliability coefficients and 36 level-2 reliability coefficients were generated, along with their corresponding standard errors, confidence intervals, and relative bias.

Multilevel Reliability Estimates Across Measurement Frameworks

When data were generated, level-1 reliability coefficients were fixed at .70 across Cronbach's α, polychoric ordinal α, and person reliability and ICC was fixed at .20, therefore, Spearman-Brown's prophecy coefficient, which uses the number of respondents at level-1 to calculate the coefficient, is fixed at .882 when $N = 30$ and fixed at .926 when $N = 50$.

Reliability and standard errors. Three noteworthy results emerged when comparing Cronbach's α and polychoric ordinal α versus person reliability at level-1 of a two-level model (reliability within level). First, Cronbach's α and polychoric ordinal α provide similar reliability estimates and standard errors across data distributions as well as across cluster and cluster sample sizes; however, person reliability overestimated reliability under conditions of normally distributed data across these same conditions. Second, person reliability was far more accurate under conditions of mixed and non-normal distributions across cluster and cluster sample sizes. Third, the standard errors of person reliability estimates were similar to those found in Cronbach's α and polychoric ordinal α across all data conditions. Table 22 shows the average reliability coefficients, standard errors, and 95% confidence intervals for each set.

Table 22

Results of Multilevel Sample Reliability Coefficients

Sample Size	Average Sample Reliability (Level 1: Within)*	Average Sample Reliability SE	Average Lower Level	Average Upper Level	Average Spearman-Brown Coefficient (Level-2: Between)	Average Spearman-Brown Coefficient SE	Average Lower Level Spearman-Brown	Average Upper Level Spearman-Brown
Cronbach's a								
All Data Normally Distributed								
Group Size 10								
30	.671	.028	.670	.672	.864	.012	.841	.888
50	.670	.026	.669	.670	.892	.009	.875	.909
Group Size 100								
30	.661	.010	.661	.661	.872	.007	.859	.885
50	.658	.009	.658	.658	.922	.003	.917	.927
Mixed Data Distribution								
Group Size 10								
30	.670	.028	.669	.671	.866	.014	.837	.894
50	.670	.022	.669	.670	.920	.005	.910	.930

Table 22 (continued)

Sample Size	Average Sample Reliability (Level 1: Within)*	Average Sample Reliability SE	Average Lower Level	Average Upper Level	Average Spearman-Brown Coefficient (Level-2: Between)	Average Spearman-Brown Coefficient SE	Average Lower Level Spearman-Brown	Average Upper Level Spearman-Brown
	Group Size 100							
30	.660	.010	.660	.661	.872	.007	.859	.885
50	.669	.023	.669	.670	.928	.005	.918	.938
	Non-Normal Distribution							
	Group Size 10							
30	.671	.028	.670	.672	.866	.014	.839	.893
50	.668	.023	.668	.669	.920	.005	.909	.930
	Group Size 100							
30	.660	.017	.660	.661	.872	.007	.859	.885
50	.658	.009	.658	.658	.922	.003	.917	.927

Table 22 (continued)

Sample Size	Average Sample Reliability (Level 1: Within)*	Average Sample Reliability SE	Average Lower Level	Average Upper Level	Average Spearman-Brown Coefficient (Level-2: Between)	Average Spearman-Brown Coefficient SE	Average Lower Level Spearman-Brown	Average Upper Level Spearman-Brown
Polychoric Ordinal a								
All Data Normally Distributed								
Group Size 10								
30	.733	.031	.732	.734	.888	.020	.848	.929
50	.704	.030	.703	.705	.929	.024	.881	.977
Group Size 100								
30	.745	.023	.744	.745	.896	.007	.882	.910
50	.699	.023	.698	.699	.927	.013	.900	.953
Mixed Data Distribution								
Group Size 10								
30	.659	.065	.654	.663	.825	.017	.791	.859
50	.659	.065	.654	.663	.825	.017	.791	.859

Table 22 (continued)

Sample Size	Average Sample Reliability (Level 1: Within)*	Average Sample Reliability SE	Average Lower Level	Average Upper Level	Average Spearman-Brown Coefficient (Level-2: Between)	Average Spearman-Brown Coefficient SE	Average Lower Level Spearman-Brown	Average Upper Level Spearman-Brown
	Group Size 100							
30	.646	.027	.645	.647	.806	.058	.690	.923
50	.746	.023	.745	.746	.935	.015	.906	.964
	Non-Normal Distribution							
	Group Size 10							
30	.680	.026	.679	.681	.845	.014	.818	.872
50	.606	.023	.605	.606	.901	.013	.876	.927
	Group Size 100							
30	.670	.010	.670	.670	.853	.008	.837	.870
50	.728	.008	.728	.728	.931	.036	.859	1.003

Table 22 (continued)

Sample Size	Average Sample Reliability (Level 1: Within)*	Average Sample Reliability SE	Average Lower Level	Average Upper Level	Average Spearman-Brown Coefficient (Level-2: Between)	Average Spearman-Brown Coefficient SE	Average Lower Level Spearman-Brown	Average Upper Level Spearman-Brown
Person Reliability (Rating Scale Model)								
All Data Normally Distributed								
Group Size 10								
30	.852	.028	.851	.853	.865	.015	.836	.895
50	.739	.023	.738	.739	.920	.005	.910	.930
Group Size 100								
30	.872	.028	.871	.873	.893	.010	.872	.913
50	.728	.009	.728	.728	.922	.003	.917	.927
Mixed Data Distribution								
Group Size 10								
30	.812	.028	.811	.813	.898	.010	.879	.917
50	.832	.029	.831	.833	.924	.020	.884	.965

Table 22 (continued)

Sample Size	Average Sample Reliability (Level 1: Within)*	Average Sample Reliability SE	Average Lower Level	Average Upper Level	Average Spearman-Brown Coefficient (Level-2: Between)	Average Spearman-Brown Coefficient SE	Average Lower Level Spearman-Brown	Average Upper Level Spearman-Brown
	Group Size 100							
30	.801	.010	.800	.801	.886	.006	.875	.897
50	.722	.061	.722	.722	.915	.010	.894	.935
	Non-Normal Distribution							
	Group Size 10							
30	.722	.028	.721	.722	.880	.009	.863	.898
50	.718	.023	.718	.719	.929	.003	.922	.935
	Group Size 100							
30	.710	.017	.710	.711	.886	.018	.851	.921
50	.708	.009	.708	.708	.931	.004	.923	.938

*Population reliability coefficients fixed at .70.

Data in Table 22 show that average Spearman-Brown's coefficients (reliability between-level) were similar across data distributions, level-1 and level-2 sample sizes, and types of reliability estimates. Furthermore, standard errors of reliability were not only low, but stable across all data conditions. Browne and Draper (2000) and Van der Leeden, eta al (1997) conducted Monte Carlo simulation studies to assess the role of standard errors in multilevel models. Browne and Draper found evidence that the higher the number of clusters, the lower the standard errors and more precise the measure. Van der Leeden et al.'s (1997) results showed that when assumptions of normality and large samples are not met, the standard errors tend to have a downward bias. The data generated for this dissertation and presented in Table 22 above provide additional evidence to support these findings. Figure 9 shows the average level-1 reliability coefficients across data conditions, and Figure 10 shows the Spearman-Brown coefficients across data conditions and type of reliability at level-1.

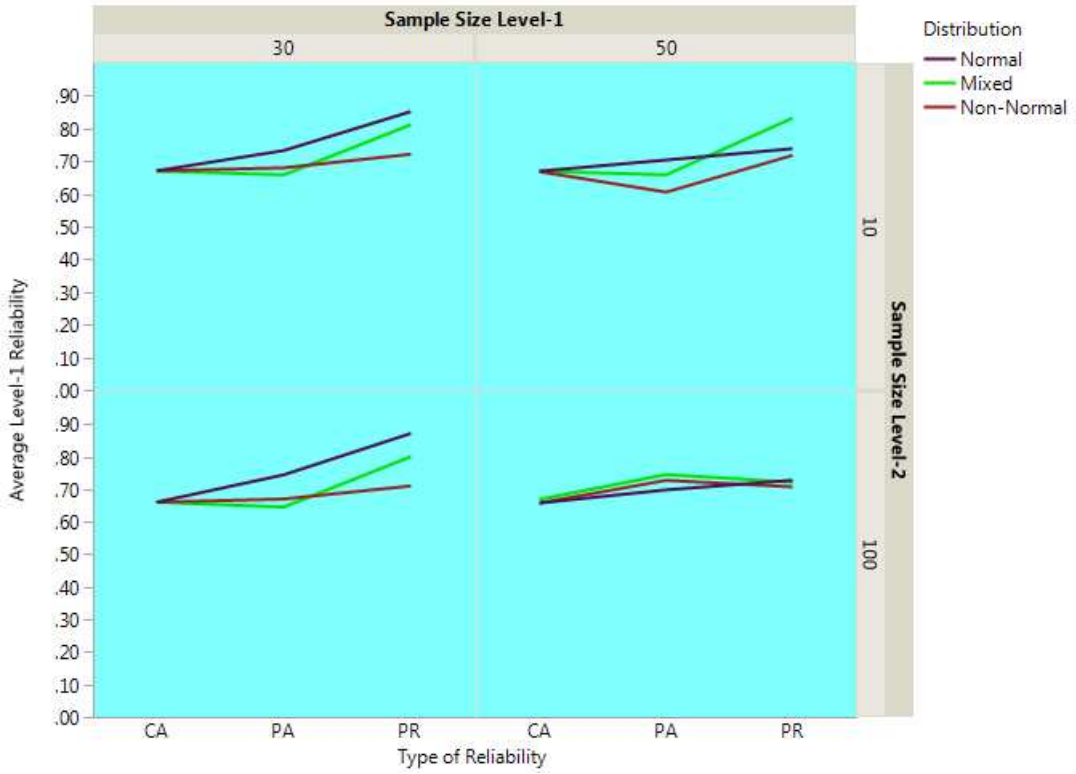

Figure 9. Average reliability coefficients at level-1 of a two-level model across data conditions.

Figures 9 and 10 provide a good visual representation that the average reliability coefficients at both the within level (Figure 9: level-1) and between-levels (Figure 10: level-2) are well within acceptable range and stable across level-1 and level-2 sample sizes, types of reliability coefficients, and types of distributions.

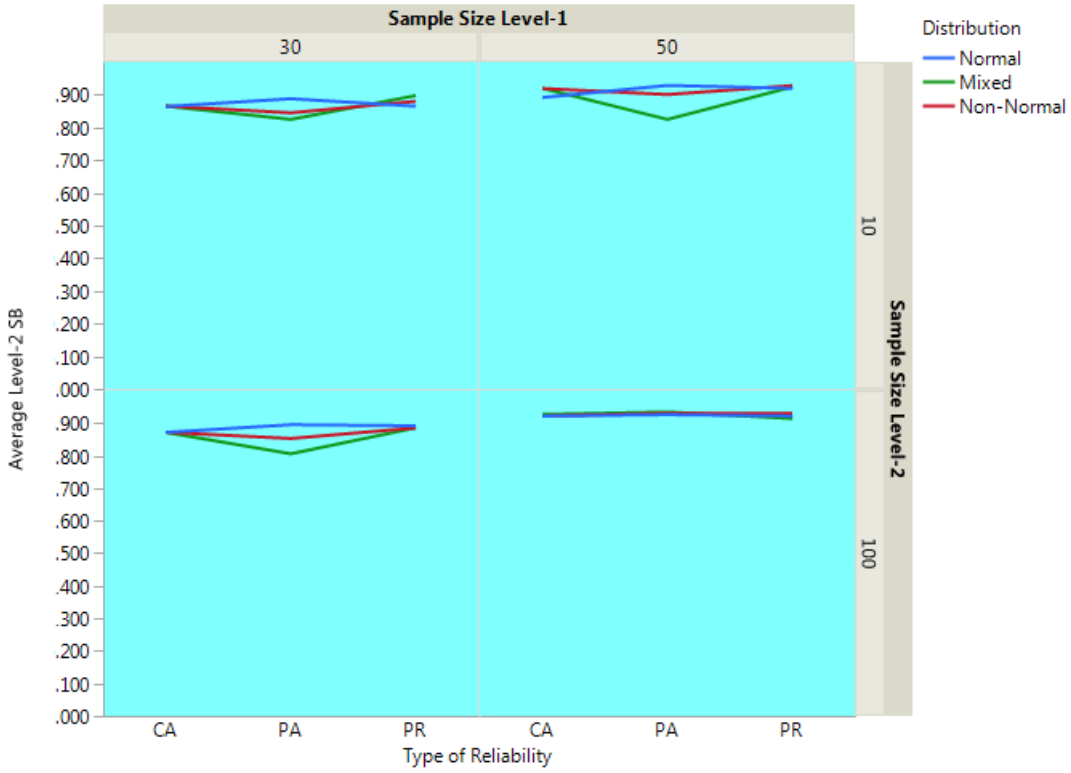

Figure 10. Average Spearman-Brown coefficients across data conditions.

Figures 11 and 12 show the average standard errors of reliability estimates in level-1 of a two-level model across data distributions, types of reliability, and cluster and cluster sample sizes. Figure 11 shows that across data conditions, standard errors at level-1 of a two-level model are quite low.

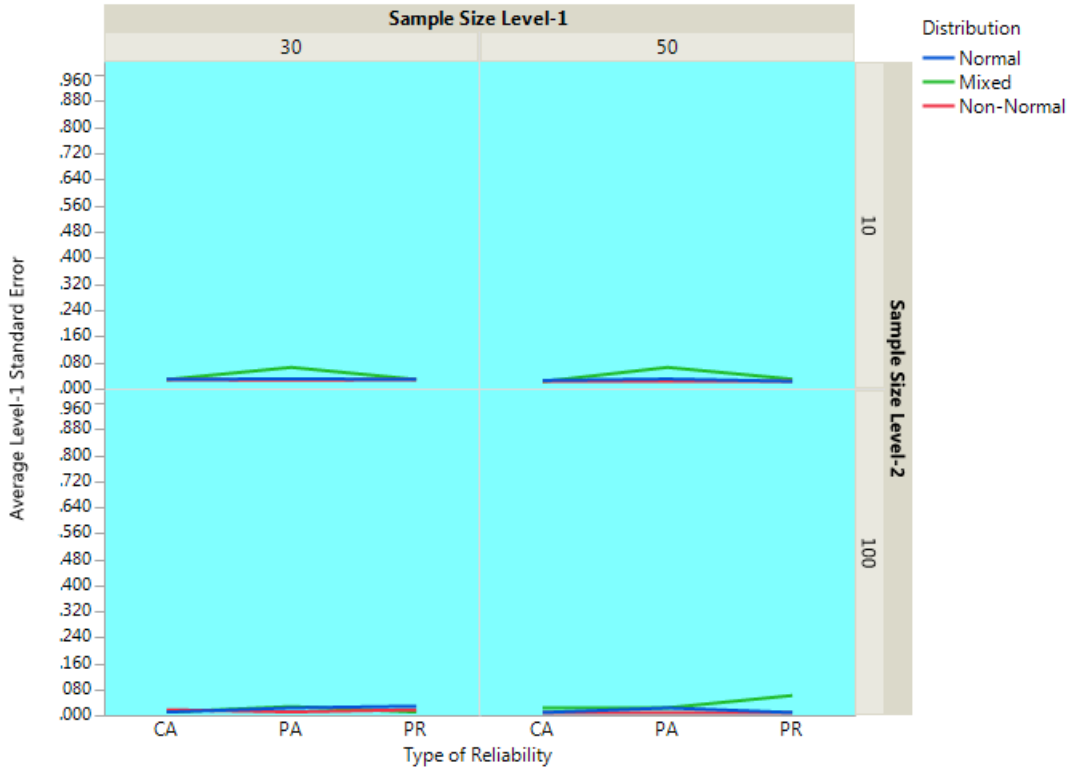

Figure 11. Standard errors in level-1 of a two-level model across data conditions.

Figure 12 shows that across data conditions, standard errors at level-2 of a two-level model are quite low and stable. The standard errors of reliability for the within and between levels across all data conditions are not only quite small, but comparable. Only under mixed distributions are any spikes (albeit small) in standard errors identified.

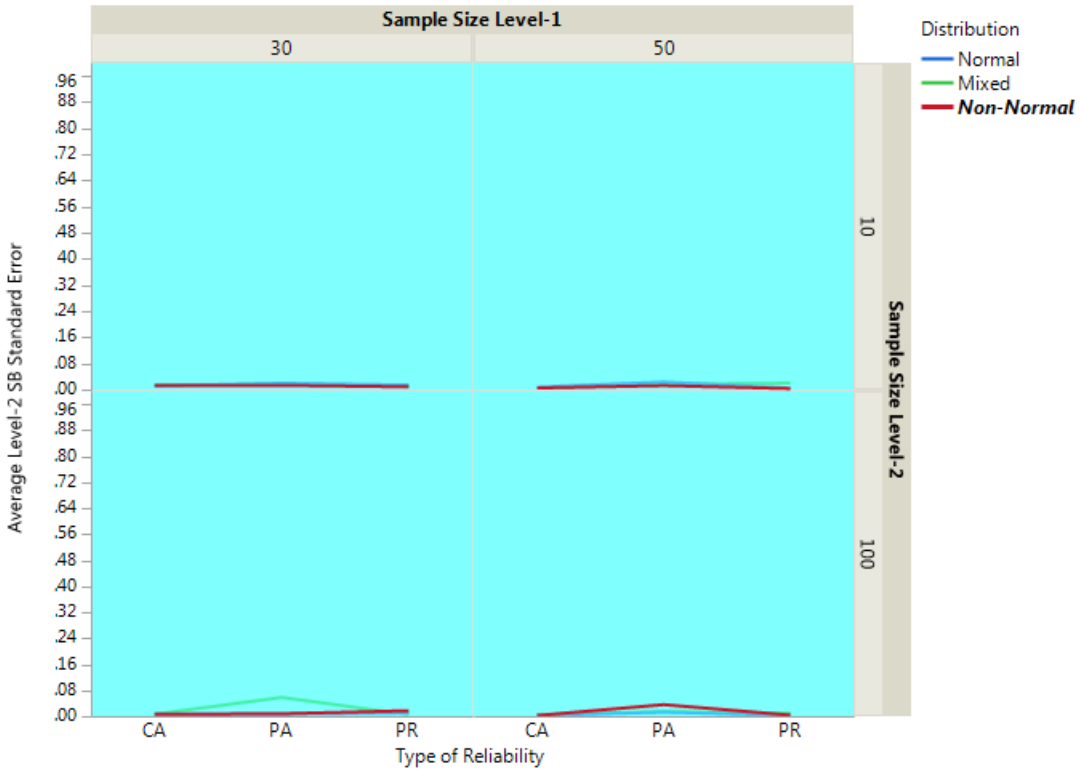

Figure 12. Level-2 standard errors across data conditions.

Reliability bias. In Table 23, the average bias in a two-level sampling design across types of reliability, data distributions, and level-1 and level-2 sample sizes is presented and four key findings are discussed below.

Table 23

Percentage of Bias in Multilevel Models for Both Level-1 and Level-2 Across Types of Reliability, Sample Sizes, and Distributions

Sample Size	Average Reliability Coefficient Relative Bias	Percentage of Reliability Coefficient Bias ≥ 10%	If Bias ≥ 10%, Percentage Underestimated	If Bias ≥ 10%, Percentage Overestimated	Average Relative Spearman-Brown (SB) Bias	Percentage of SB Bias ≥ 10%	If Bias ≥ 10%, Percentage Underestimated	If Bias ≥ 10%, Percentage Overestimated
Cronbach's α								
All Data Normally Distributed								
Group Size 10								
30	.042	24.20%	24.20%	0.00%	.020	0.40%	0.40%	0.00%
50	.045	20.40%	20.40%	0.00%	.007	0.00%	0.00%	0.00%
Group Size 100								
30	.056	16.30%	16.30%	0.00%	.012	0.00%	0.00%	0.00%
50	.060	21.40%	21.40%	0.00%	.004	0.00%	0.00%	0.00%

Table 23 (continued)

Sample Size	Average Reliability Coefficient Relative Bias	Percentage of Reliability Coefficient Bias ≥ 10%	If Bias ≥ 10%, Percentage Underestimated	If Bias ≥ 10%, Percentage Overestimated	Average Relative Spearman-Brown (SB) Bias	Percentage of SB Bias ≥ 10%	If Bias ≥ 10%, Percentage Underestimated	If Bias ≥ 10%, Percentage Overestimated
Mixed Data Distribution								
Group Size 10								
30	.043	24.30%	24.30%	0.00%	.019	0.40%	0.00%	0.00%
50	.043	19.40%	19.40%	0.00%	.007	0.00%	0.00%	0.00%
Group Size 100								
30	.056	19.00%	19.00%	0.00%	.012	0.00%	0.00%	0.00%
50	.044	22.80%	22.80%	0.00%	.002	0.00%	0.00%	0.00%
Non-Normal Distribution								
Group Size 10								
30	.042	21.60%	21.50%	0.50%	.019	0.00%	0.00%	0.00%
50	.045	21.30%	21.30%	0.00%	.007	0.00%	0.00%	0.00%

Table 23 (continued)

Sample Size	Average Reliability Coefficient Relative Bias	Percentage of Reliability Coefficient Bias ≥ 10%	If Bias ≥ 10%, Percentage Underestimated	If Bias ≥ 10%, Percentage Overestimated	Average Relative Spearman-Brown (SB) Bias	Percentage of SB Bias ≥ 10%	If Bias ≥ 10%, Percentage Underestimated	If Bias ≥ 10%, Percentage Overestimated
Group Size 100								
30	.057	17.30%	17.30%	0.00%	.011	0.00%	0.00%	0.00%
50	.060	20.50%	20.50%	0.00%	.004	0.00%	0.00%	0.00%
Polychoric Ordinal a								
All Data Normally Distributed								
Group Size 10								
30	.051	27.60%	0.19%	27.41%	.007	0.00%	0.00%	0.00%
50	.026	8.60%	3.90%	4.70%	.003	0.00%	0.00%	0.00%
Group Size 100								
30	.066	45.90%	0.02%	45.88%	.015	0.00%	0.00%	0.00%
50	.026	3.80%	2.79%	1.01%	.000	0.00%	0.00%	0.00%

Table 23 (continued)

Sample Size	Average Reliability Coefficient Relative Bias	Percentage of Reliability Coefficient Bias ≥ 10%	If Bias ≥ 10%, Percentage Underestimated	If Bias ≥ 10%, Percentage Overestimated	Average Relative Spearman-Brown (SB) Bias	Percentage of SB Bias ≥ 10%	If Bias ≥ 10%, Percentage Underestimated	If Bias ≥ 10%, Percentage Overestimated
Mixed Data Distribution								
Group Size 10								
30	.053	35.90%	35.90%	0.00%	.065	8.40%	8.40%	0.00%
50	.043	27.00%	27.00%	0.00%	.004	0.00%	0.00%	0.00%
Group Size 100								
30	.077	58.30%	58.30%	0.00%	.086	46.80%	46.80%	0.00%
50	.065	45.30%	0.00%	45.30%	.010	0.00%	0.00%	0.00%
Non-Normal Distribution								
Group Size 10								
30	.042	14.00%	13.70%	0.30%	.042	0.00%	0.00%	0.00%
50	.045	98.40%	98.40%	0.00%	.027	0.00%	0.00%	0.00%

Table 23 (continued)

Sample Size	Average Reliability Coefficient Relative Bias	Percentage of Reliability Coefficient Bias ≥ 10%	If Bias ≥ 10%, Percentage Underestimated	If Bias ≥ 10%, Percentage Overestimated	Average Relative Spearman-Brown (SB) Bias	Percentage of SB Bias ≥ 10%	If Bias ≥ 10%, Percentage Underestimated	If Bias ≥ 10%, Percentage Overestimated
Group Size 100								
30	.057	3.80%	3.80%	0.00%	.033	0.00%	0.00%	0.00%
50	.060	0.30%	0.00%	0.30%	.005	0.00%	0.00%	0.00%
Person Reliability (Rating Scale Model)								
All Data Normally Distributed								
Group Size 10								
30	.217	99.80%	0.00%	99.80%	.019	0.50%	0.50%	0.00%
50	.055	34.00%	0.00%	34.00%	.007	0.00%	0.00%	0.00%
Group Size 100								
30	.042	99.90%	0.00%	99.90%	.012	0.00%	0.00%	0.00%
50	.043	10.00%	0.00%	10.00%	.004	0.00%	0.00%	0.00%

Table 23 (continued)

Sample Size	Average Reliability Coefficient Relative Bias	Percentage of Reliability Coefficient Bias ≥ 10%	If Bias ≥ 10%, Percentage Underestimated	If Bias ≥ 10%, Percentage Overestimated	Average Relative Spearman-Brown (SB) Bias	Percentage of SB Bias ≥ 10%	If Bias ≥ 10%, Percentage Underestimated	If Bias ≥ 10%, Percentage Overestimated
Mixed Data Distribution								
Group Size 10								
30	.056	96.20%	0.00%	96.20%	.018	0.00%	0.00%	0.00%
50	.044	99.30%	0.00%	100.00%	.002	0.00%	0.00%	0.00%
Group Size 100								
30	.020	100.00%	0.00%	100.00%	.004	0.00%	0.00%	0.00%
50	.032	18.00%	8.33%	91.67%	.012	0.00%	0.00%	0.00%
Non-Normal Distribution								
Group Size 10								
30	.031	17.90%	1.80%	16.10%	.000	0.00%	0.00%	0.00%
50	.026	2.80%	0.20%	2.60%	.001	0.00%	0.00%	0.00%

Table 23 (continued)

Sample Size	Average Reliability Coefficient Relative Bias	Percentage of Reliability Coefficient Bias ≥ 10%	If Bias ≥ 10%, Percentage Underestimated	If Bias ≥ 10%, Percentage Overestimated	Average Relative Spearman-Brown (SB) Bias	Percentage of SB Bias ≥ 10%	If Bias ≥ 10%, Percentage Underestimated	If Bias ≥ 10%, Percentage Overestimated
Group Size 100								
30	.017	3.20%	2.20%	1.00%	.046	0.00%	0.00%	0.00%
50	.011	3.00%	2.10%	0.90%	.050	0.00%	0.00%	0.00%

First, average relative bias in reliability estimates at both level-1 (within clusters) and level 2 (between clusters) is less than 10% across data distributions, cluster and cluster sample sizes, and types of reliability coefficients with the exception of person reliability under the conditions of small cluster and cluster sample sizes and normally distributed data. This finding was unexpected and, along with the results of person reliability in single-level sampling designs, is discussed in detail in Chapter V. Second, since relative bias was calculated for every reliability coefficient, if the coefficient had bias ≥ 10%, it was counted and included in the calculation for percentage of bias ≥ 10% and the direction of bias was noted. Cronbach's α underestimated the reliability coefficient under every data condition while polychoric ordinal α overestimated the reliability coefficient across cluster and cluster sample sizes under normal and mixed data distributions, with one exception (when cluster N = 50 and cluster N = 100), and underestimated reliability under non-normal data distributions. With very few exceptions, person reliability overestimated reliability coefficients almost every time. Third, Spearman-Brown coefficients, calculated at the between clusters level showed very little, if any, bias and were a stable indicator of between-level reliability across all data conditions with the exception of polychoric ordinal α under a mixed distribution when the number of clusters is 100 and the number of clusters is 30.

Tests of Hypotheses for Research Question 2

Three factorial ANOVAs were conducted with dependent variables split between level-1 and level-2 for ease of analysis. Therefore, six F statistics are reported, one for each dependent variable (reliability coefficient for level-1 and level-2, standard errors of measurement for levels-1 and -2 combined, percentage of relative bias for level-1 and

level-2), and one Chi-square test for direction of reliability estimates with ≥ 10% bias (low or high), with three level-1 reliability coefficients (Cronbach's α, polychoric ordinal α, and person reliability), and one level-2 reliability coefficient (Spearman-Brown), across two level-1 sample sizes (N = 30 and N = 50) and two level-2 sample sizes (N =10 and N = 100), three data distributions (normal, mixed, and non-normal), and two measurement frameworks (CTT and RSM). Table 24 shows the results of the factorial ANOVA for level-1 (within level) reliability coefficients across sample sizes and distributions.

Table 25 represents the assessment of the level-1 reliability coefficients when the interaction plots failed to show an interaction (e.g. the lines were not crossed). Note that the interaction between type of reliability and type of distribution, and the level-1 sample size and type of distribution were statistically significant; however, the effect sizes are < .06, which I translated as a small magnitude of difference.

Table 26 shows the results of the factorial ANOVA for Level-2 (between-level) reliability coefficients across level-1 and level-2 sample sizes and types of distributions. All results are statistically significant and therefore, effect sizes were calculated for every variable and all interactions.

Table 27 shows the assessment of simple main effects for level-2 reliability coefficients across the type of distribution and sample size when interactions were present. Most notably, level-1 and level-2 sample sizes and types of distributions have a large effect on level-2 reliability coefficients.

Table 28 shows the level-1 standard errors of reliability across types of distribution and sample sizes. All interactions are statistically, but not substantially significant, based on effect sizes. Simple main effects were assessed and reported below.

Table 24

Results for Factorial ANOVA for Level-1 Reliability Coefficients Across Sample Size and Distribution

Source	*df*	*F**	Effect Size □2
Distribution	2	4815.504	.061
Type of Reliability	2	37710.313	.476
Level-1 Sample Size	1	1416.941	.009
Level-2 Sample Size	1	573.656	.003
Distribution * Type of Reliability	2	4469.949	.009
Distribution * Level-1 Sample Size	2	2465.146	.031
Distribution * level-2 Sample Size	2	527.178	.006
Type of Reliability * Level-1 Sample Size	2	5881.11	.056
Type of Reliability * Level-2 Sample Size	2	637.202	.008
Level-1 Sample Size * Level 2 Sample Size	1	20.109	.000
Distribution * Type of Reliability * Level-1 Sample Size	2	2293.846	.029
Distribution * Type of Reliability * Level-2 Sample Size	2	744.493	.009
Distribution * Level-1 Sample Size* Level-2 Sample Size	2	278.815	.003
Type of Reliability * Level-1 Sample Size * Level-2 Sample Size	2	1250.841	.015
Distribution * Type of Reliability * Level-1 Sample Size * Level-2 Sample Size	2	422.565	.005

*$p < .0001$

Table 25

Assessment of Level-1 Reliability Coefficients When Interaction Plots Showed No Interaction

Source	Test Statistic	*p*-value	Post-hoc Analysis/Explanation	Eta Squared
Type of Reliability Coefficient * Distribution	*F = 4469.949*	$p < .0001$	Person reliability coefficients were higher than Cronbach's α or polychoric ordinal α across distributions as well as higher under conditions of normally distributed data	.056
Level-1 Sample Size * Distribution	*F = 2465.146*	$p < .0001$	Reliability coefficients were higher when level-1 sample size $N = 30$ across all distributions, with normal and mixed distributions showing the highest reliability coefficients	.031
Level-1 Sample Size * Level-2 Sample Size	*F = 20.109*	$p = .131$	Reliability coefficients were higher when level-2 sample size $N = 10$ across both level-1 sample sizes, with the level-1 N =1 50 and level-2 N = 10	.00012

Table 26

Results for Factorial ANOVA for Level-2 Reliability Coefficients Across Sample Size and Distribution

Source	*df*	*F**	Eta Squared
Distribution	2	126.681	.002
Type of Reliability	2	1477.265	.019
Level-1 Sample Size	1	79940.126	.527
Level-2 Sample Size	1	333.614	.002
Distribution * Type of Reliability	2	70.454	.004
Distribution * Level-1 Sample Size	2	127.101	.002
Distribution * Level-2 Sample Size	2	254.471	.006
Type of Reliability * Level-1 Sample Size	2	2028.602	.027
Type of Reliability * Level-2 Sample Size	2	75.443	.000
Level-1 Sample Size * Level 2 sample Size	1	0.154	.000
Distribution * Type of Reliability * Level-1 Sample Size	2	241.415	.003
Distribution * Type of Reliability * Level-2 Sample Size	2	362.072	.005

Table 26 (continued)

Source	*df*	*F**	Eta Squared
Distribution * Level-1 Sample Size* Level-2 Sample Size	2	135.684	.001
Type of Reliability * Level-1 Sample Size * Level-2 Sample Size	2	424.075	.006
Distribution * Type of Reliability * Level-1 Sample Size * Level-2 Sample Size	2	422.565	.000

* $p < .0001$

Table 27

Level-2 Reliability Coefficients' Across Distribution and Sample Size: Assessment of Simple Main Effects

Source	*F**	Post-hoc Analysis/Explanation	Eta Squared
Reliability Type * Distribution	$F = 957.749$	Person reliability had the highest level-2 reliability coefficients across distributions, with the non-normal distribution showing the highest level-2 person reliability coefficients. Cronbach's a and polychoric ordinal a had stable level-2 reliability estimates across distributions	.026
Level-1 Sample Size * Distribution	$F = 44334.847$	Level-2 reliability coefficients were higher for level-1 $N = 100$, than for level-1 $N = 10$ across distributions.	.346
Level-1 Sample Size * Type of Reliability	$F = 31200.974$	Level-2 reliability coefficients were higher for level-1 $N = 50$, than for level-1 $N = 30$ across types of reliability	.456
Level-2 Sample Size * Type of Reliability	$F = 335.912$	Level-2 reliability coefficients were higher for level-2 $N = 100$ than for level-2 $N = 10$ across types of reliability.	.009

*$p < .0001$

Table 28

Level-1 Standard Errors of Reliability Across Distribution and Sample Sizes

Source	*df*	*F**	*p*-value	Effect Size □2
Distribution	2	1808.616	$p < .0001$	.035
Reliability Type	2	4613.427	$p < .0001$	.091
Level-1 Sample Size	1	855.502	$p < .0001$	.028
Level-2 Sample Size	1	142.114	$p < .0001$	.025
Distribution * Type of Reliability	2	30.621	$p < .0001$	.000
Distribution * Level-1 Sample Size	2	517	$p < .0001$	.002
Distribution * Level-2 Sample Size	2	143.76	$p < .0001$	.003
Type of Reliability * Level-1 Sample Size	2	335.303	$p < .0001$	.043
Type of Reliability * Level-2 Sample Size	2	435.671	$p < .0001$	.002
Level-1 Sample Size * Level 2 sample Size	1	44.006	$p < .0001$	.004
Distribution * Type of Reliability * Level-1 Sample Size	2	470.351	$p < .0001$	.000
Distribution * Type of Reliability * Level-2 Sample Size	2	221.049	$p < .0001$	.000

Table 28 (continued)

Source	*df*	*F**	*p*-value	Effect Size □2
Distribution * Level-1 Sample Size* Level-2 Sample Size	1	26.371	$p < .0001$	.003
Type of Reliability * Level-1 Sample Size * Level-2 Sample Size	3	274.891	$p < .0001$	.000
Distribution * Type of Reliability * Level-1 Sample Size * Level-2 Sample Size	1	64.922	$p < .0001$	.000

* $p < .0001$

Table 29

Level-1 Standard Errors of Reliability Coefficients: Main Effects

Source	*F**	Post-hoc Analysis/Explanation	eta squared
Type of Reliability * Distribution	*F* = 816.265	Person standard errors are higher than Cronbach's α or polychoric ordinal α standard errors under mixed or non-normal distributions	.001
Level-2 Sample Size * Type of Reliability	*F* = 435.671	Polychoric ordinal α has higher standard errors when level-2 sample size *N* = 100 than Cronbach's α or person reliability	.015
Level-2 Sample Size * Distribution	*F*= 143.76	When level-2 sample size *N* = 100, standard errors are smaller than when level-2 sample size *N* = 10	.005

**p* <.0001

Table 30

Level-1 Relative Bias ≥ 10% Across Distribution and Sample Sizes

Source	*df*	*F*	*p*-value	Effect Size □2
Distribution	2	.849	$p = .428$	.000
Type of Reliability	2	1391.822	$p < .0001$	.059
Level-1 Sample Size	1	.024	$p = .877$	.000
Level-2 Sample Size	1	140.948	$p < .0001$	.003
Distribution * Type of Reliability	2	.955	$p = .385$	.000
Distribution * Level-1 Sample Size	2	.045	$p = .956$	.000
Distribution * Level-2 Sample Size	2	7.865	$p = .155$	.000
Type of Reliability * Level-1 Sample Size	2	1.500	$p = .223$	.000
Type of Reliability * Level-2 Sample Size	2	101.485	$p < .0001$	.005
Level-1 Sample Size * Level 2 Sample Size	1	62.908	$p < .0001$	.002
Distribution * Type of Reliability * Level-1 Sample Size	2	1.435	$p = .223$	.000
Distribution * Type of Reliability * Level-2 Sample Size	2	.589	$p = .555$	.000
Distribution * Level-1 Sample Size* Level-2 Sample Size	1	1.502	$p = .223$	.000
Type of Reliability * Level-1 Sample Size * Level-2 Sample Size	1	7.221	$p = .07$	.013

Table 31

Count of Level-1 Relative Bias ≥ 10% Across All Data Conditions

Sample Size	If Bias ≥ 10%, Count Underestimated	If Bias ≥ 10%, Count Overestimated
Cronbach's α		
All Data Normally Distributed		
Group Size 10		
30	242	0
50	204	0
Group Size 100		
30	163	0
50	214	0
Mixed Data Distribution		
Group Size 10		
30	243	0
50	194	0
Group Size 100		
30	190	0
50	228	0
Non-Normal Distribution		
Group Size 10		
30	215	5
50	213	0

Table 31 (continued)

Sample Size	If Bias ≥ 10%, Count Underestimated	If Bias ≥ 10%, Count Overestimated
Group Size 100		
30	173	0
50	205	0
Polychoric ordinal α		
All Data Normally Distributed		
Group Size 10		
30	0	274
50	37	47
Group Size 100		
30	0	459
50	3	10
Mixed Data Distribution		
Group Size 10		
30	359	0
50	270	0
Group Size 100		
30	583	0
50	0	453

Table 31 (continued)

Sample Size	If Bias ≥ 10%, Count Underestimated	If Bias ≥ 10%, Count Overestimated
Non-Normal Distribution		
Group Size 10		
30	137	3
50	984	0
Group Size 100		
30	38	0
50	0	3
Person Reliability (Rating Scale Model)		
All Data Normally Distributed		
Group Size 10		
30	0	998
50	0	340
Group Size 100		
30	0	999
50	0	100
Mixed Data Distribution		
Group Size 10		
30	0	962
50	0	1000
Group Size 100		
30	0	1000
50	0	917

Table 31 (continued)

Sample Size	If Bias ≥ 10%, Count Underestimated	If Bias ≥ 10%, Count Overestimated
Non-Normal Distribution		
Group Size 10		
30	18	161
50	2	26
Group Size 100		
30	0	0
50	0	0

Table 32

Level-1 Direction of Relative Bias ≥ 10% Across Distribution and Sample Sizes (Chi-square Results)

Source	*df*	*Chi-square*	*p*-value
Pearson Chi-square			
Distribution	2	6279.418	$p < .0001$
Type of Reliability	2	12119.805	$p < .0001$
Level-1 Sample Size	1	502.163	$p < .0001$
Level-2 Sample Size	1	2.098	$p = .147$

Level-1 Reliability Coefficients as the Dependent Variable

Significant interactions were found in the factorial ANOVAs used to answer research question 2, and effect sizes are included in the analysis below. The analysis of a factorial ANOVA requires the assessment of interactions rather than examining the main

effects first. A significant interaction effect means that the effect of one independent variable depends on the value, or level, of some other independent variable included in the study design (Jaccard & Turrisi, 2003; Oshima & MCcarty, 2015). For example, as seen in Table 24 presented above, the effect of level-1 sample size depends on the type of data distribution as well as the type of reliability. When significant interaction effects are found, caution must be taken when interpreting the results of any corresponding main effects. Interpreting the main effects when significant interaction effects are present can lead to invalid conclusions. One strategy recommended by Jaccard and Turrisi (2003) and Oshima and MCcarty (2015) in a study design with two independent variables is to assign one independent variable as a focal variable and another as a moderator variable. Another strategy is to examine the graph of cell means and conduct tests of simple main effects, holding one independent variable constant while assessing the effect of another.

Since these data provide evidence of significant one, two, three, and four-way interactions with $\alpha = .05$, I chose first to graph the marginal cell means, which are the means from one independent variable averaged across all levels of another independent variable, and report interactions only for those graphs showing interactions, second, calculate effect sizes for all significant interactions and following Cohen's (1977, 1988) rule of thumb for medium effect sizes and report only interactions with effect sizes > .06 in the analysis below, third assess the simple main effects results for interactions with effect sizes > .06, and fourth, assess the results of simple main effects where the interaction found in the factorial ANOVA was significant but plotted interactions were not present.

For level-1 of the two-level models, interaction plots did not illustrate interaction effects for (a) the type of reliability * distribution, (b) level-1 sample size * distribution, and (c) level-1 sample size * level-2 sample size; therefore, main effects are reported in Table 25. Figures 13 to 15 represent the graphic displays of level-1 reliability coefficient interaction effects created in SPSS version 24.0 and the simple main effects are presented following each significant interaction plot.

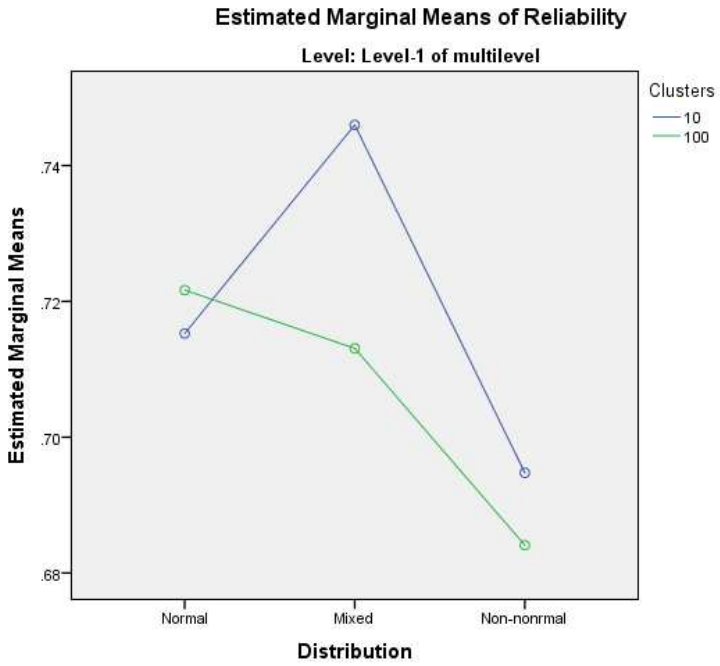

Figure 13. Interaction effects where the marginal means of reliability estimates based on level-2 sample size are averaged across the type of data distribution.

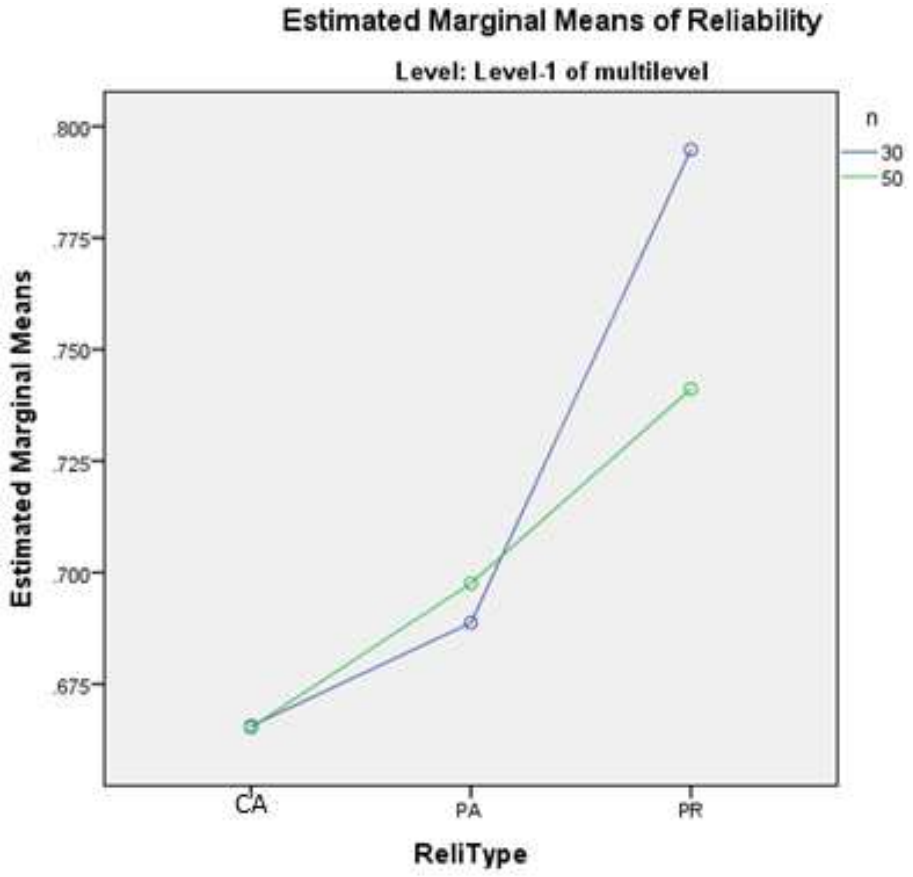

Figure 14. A graphical representation of interaction effects of level-1 sample size marginal means average across type of reliability coefficient.

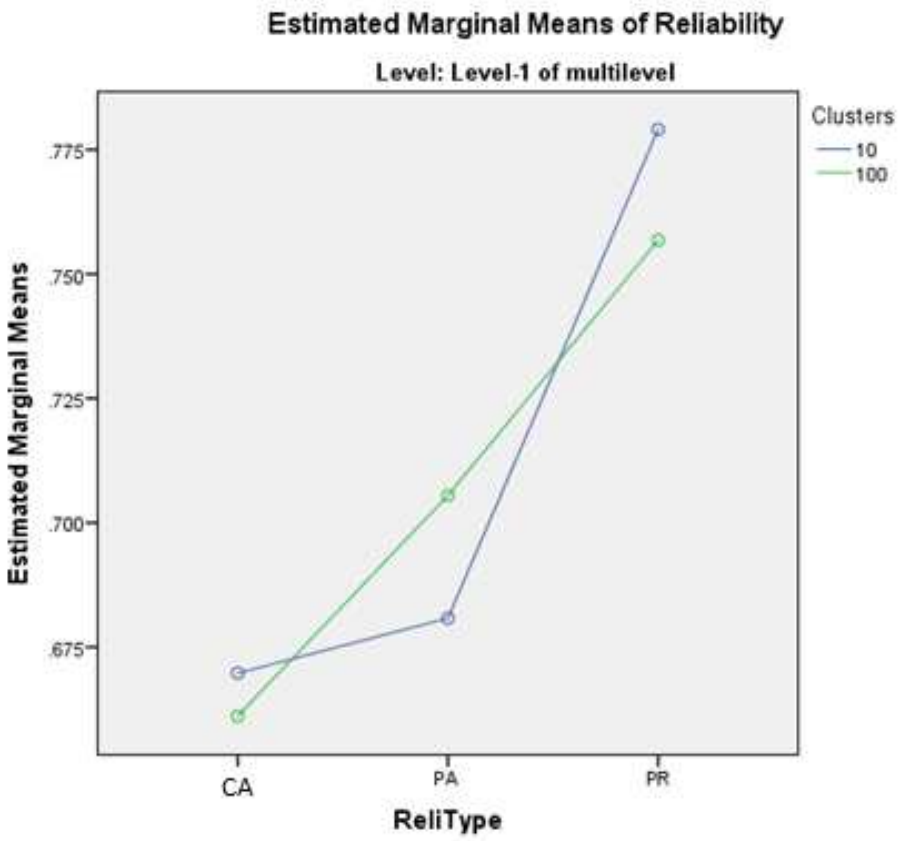

Figure 15. A graphical representation of interaction effects of level-2 sample sizes across types of reliability.

Figure 13 shows that for level-2 sample size of a two-level model, $N = 10$ led to higher reliability estimates than $N = 100$ in mixed distributions ($F[1,35972] = 2333.985$, $p < .0001$, $\eta^2 = .061$) and non-normal distributions ($F[1,35972] = 244.888$ $p < .0001$, $\eta^2 = .006$), while for normal distributions, level-2 $N = 10$ led to lower reliability coefficients than when $N = 100$ (F [1,35,972] = 170.044, $p < .0001$, $\eta^2 = .004$).

Figure 14 represents the interaction effects where the marginal means of reliability estimates based on level-1 sample size are averaged across the type of data reliability coefficient.

Figure 14 shows that person reliability led to higher reliability estimates when level-1 sample size $N = 30$ than when level- sample sizes $N = 50$, (F [1, 35,972] = 9302.968, $p < .0001$, $\eta^2 = .244$), indicating that the amount of variance in person reliability coefficients was dependent on level-1 sample size. Figure 15 represents the interaction effects where the marginal means of reliability estimates based on level-2 sample size are averaged across the type of data reliability coefficients.

Figure 15 shows that for level-2 (i.e., clusters) sample of a two-level model, $N =$ 10 led to higher Cronbach's α (F[1, 35,972] = 239.678, $p <$,0001, $\eta^2 = .006$) and person reliability estimates (F[1, 35,972] = 1603.804, $p <$,0001, $\eta^2 = .042$) than when level-2 sample size $N = 100$. Finally, sample size $N = 10$ led to lower polychoric α coefficients.

The results of interactions present in the ANOVA Table 24 but not in the interaction plots are in Table 25. Effect sizes were calculated for each interaction effect and type of reliability coefficient. Distribution had the largest effect size presented. These results show that in level-1 of a two-level sampling design, (a) person reliability was higher than Cronbach's α and polychoric ordinal α, and level-1 sample size is confounded with data distributions.

The results of these analyses reveal that when considering the size of the effect a given independent variable has on the amount of variance accounted for in the level-1 reliability coefficient, the type of distribution and the type of reliability coefficient have the greatest effect.

Level-2 Reliability Coefficients as the Dependent Variable

Table 26 shows the results of the factorial ANOVA when the dependent variable is the level-2 reliability coefficient (Spearman-Brown). The independent variables are the

type of reliability coefficient at level-1 of the two-level model, level-1 and level-2 sample sizes, and the underlying data distribution. Effect sizes are reported with the analysis below.

All interactions were statistically significant except level-1 * level-2 sample sizes; therefore, I followed the same procedures of analysis described above when the dependent variable was level-1 reliability coefficients. The plot of interactions presented no discernable interactions for (a) type of level-1 reliability coefficient * the distribution, (b) level-1 sample sizes * distribution, (c) level-2 sample sizes * type of reliability coefficient, and (d) level-1 * level-2 sample sizes. Figure 16 shows the interaction between level-2 sample size and data distribution where level-2 sample size $N = 10$ led to higher level-2 reliability coefficients under the condition of mixed distributions than when level-2 $N = 100$ ($F[1, 35971] = 26.988$, $p < .0001$, $\eta^2 = .007$). Furthermore, Spearman-Brown (level-2 reliability coefficient) is statistically significantly lower when level-2 $N = 10$ under normal distributions than under non-normal distributions ($F[1, 35971] = 966.615$ $p < .0001$, $\eta^2 = .026$); however, not substantially lower, based on the small effect size.

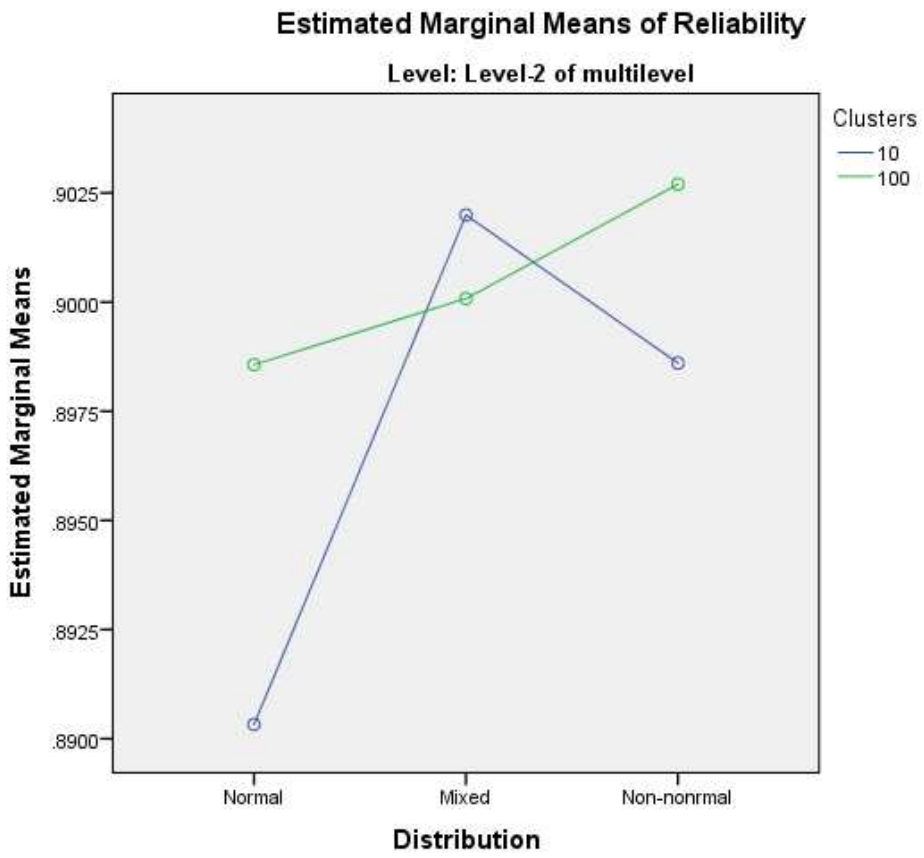

Figure 16. A graphical representation of the interaction between level-2 sample sizes and data distributions for level-2 reliability coefficients.

Table 27 shows the results of the simple main effects for (a) level-1 sample size * type of reliability, and (b) type of reliability * distribution since the interaction plots did not present interaction effects.

Assessing each of the interaction effects plots and results from the factorial ANOVA, I found that the type of reliability and the type of data distribution accounted for a substantial amount of the variance found in the level-2 Spearman-Brown coefficient. Person reliability and non-normal distributions led to higher level-2 reliability coefficients. Level-2 sample size $N = 100$ led to higher level-2 reliability coefficients than

when level-2 sample size N = 10 across distributions. Finally, level-1 N = 50 led to higher level-2 reliability coefficients than when level-1 N = 30.

Level-1 Standard Errors of Reliability Estimates

Table 28, presented above, provides the results for the factorial ANOVA where the dependent variable is the standard errors of reliability coefficients in level-1 of a two-level model and the independent variables are the type of data distribution, type of reliability coefficient, and level-1 and level-2 sample sizes. Effect sizes are reported with the analysis below.

Interactions for Standard Errors at Level-1

With α =.05, all interactions were significant. The graph of interactions presented no discernable interactions for (a) Type of Reliability * Distribution, (b) level-2 sample size * Distribution, and (c) level-2 sample size * Type of reliability; therefore, I followed the same procedures of analysis described above when the dependent variable was level-1 standard errors. Figure 17, presented previously and mentioned now as a comparison shows the interaction between level-1sample size and data distribution where level-1 sample size N = 30 led to higher level-1 standard errors under the condition of non-normal distributions than when level-1 N = 50 (F[1, 35971] = 2497.47, p < .0001, η^2 = .164) but higher standard errors when level-1 N = 50 when data were based on a mix of normal and non-normal items.

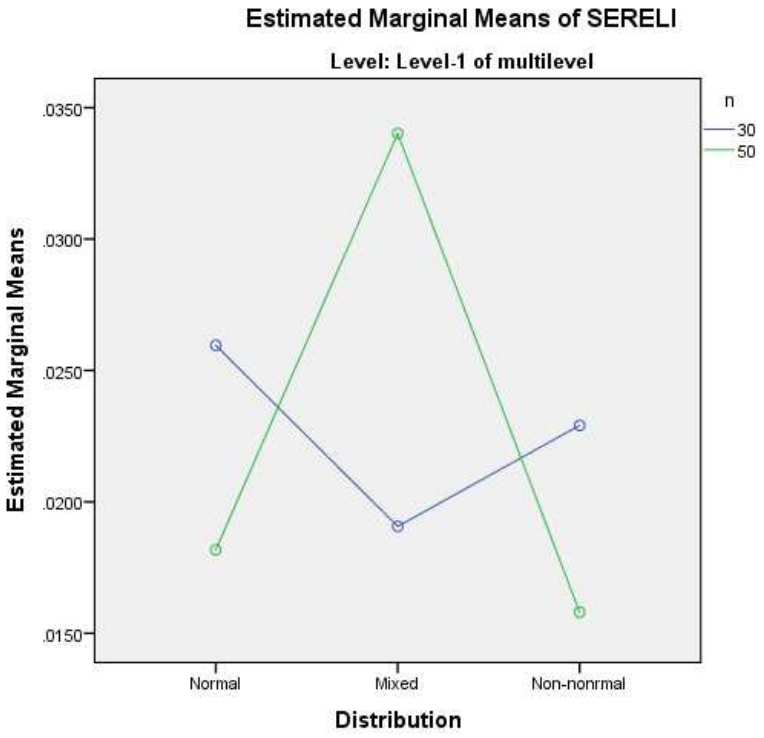

Figure 17. Interaction effect on standard errors of reliability coefficients between level-1 sample sizes and type of distribution.

Figure 18 shows the interaction effect on level-1 standard errors between level-1 sample size and type of reliability coefficient. Standard errors are lower when level-1 N = 30 for person reliability estimates than when N = 50 (F[1, 36971] = 138.131, p < .0001, $\eta^{2 =}$ .004); however, this pattern was reversed for polychoric ordinal α and Cronbach's α where reliability estimates were higher when the level-1 N = 30 (F[1, 36971] = 4082.838, p < .0001, $\eta^{2 =}$ .116).

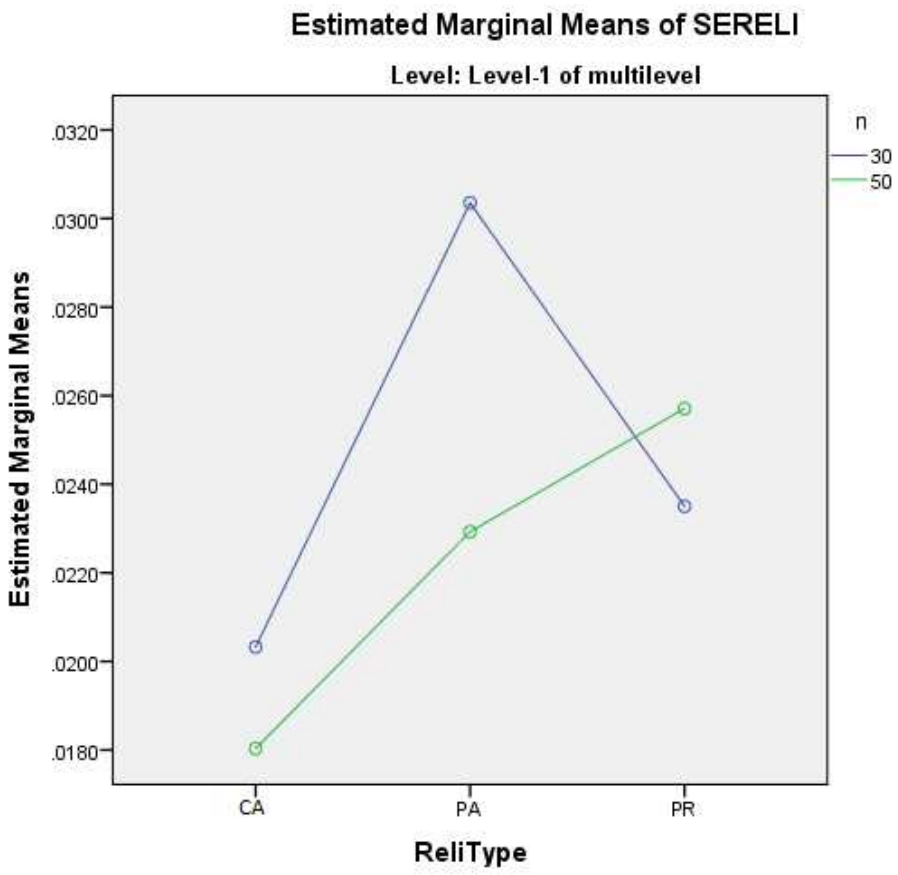

Figure 18. A graphical representation of the interaction effect on level-1 standard errors of reliability (SERELI) between level-1 sample size and type of reliability coefficient (ReliType).

These results suggest that level-1 sample size $N = 30$ leads to higher standard errors for polychoric ordinal α and lower standard errors for person reliability than when level-1 sample size $N = 50$. Figure 19 shows the dis-ordinal interaction effect between level-1 and level-2 sample sizes on standard errors. Standard errors are lower when level-1 sample size $N = 30$ and level-2 sample size $N = 100$ than when level-1 sample size = 50 and level-2 sample size $N = 10$ ($F[1, 35021)] = 3041.397$, $p < .0001$, $\eta^{2} = .082$) suggesting that standard errors were at their highest when overall N was also at its lowest with only

10 clusters based on 30 cases each, whereas standard errors were at their lowest when there were 100 clusters also based on 30 cases per cluster.

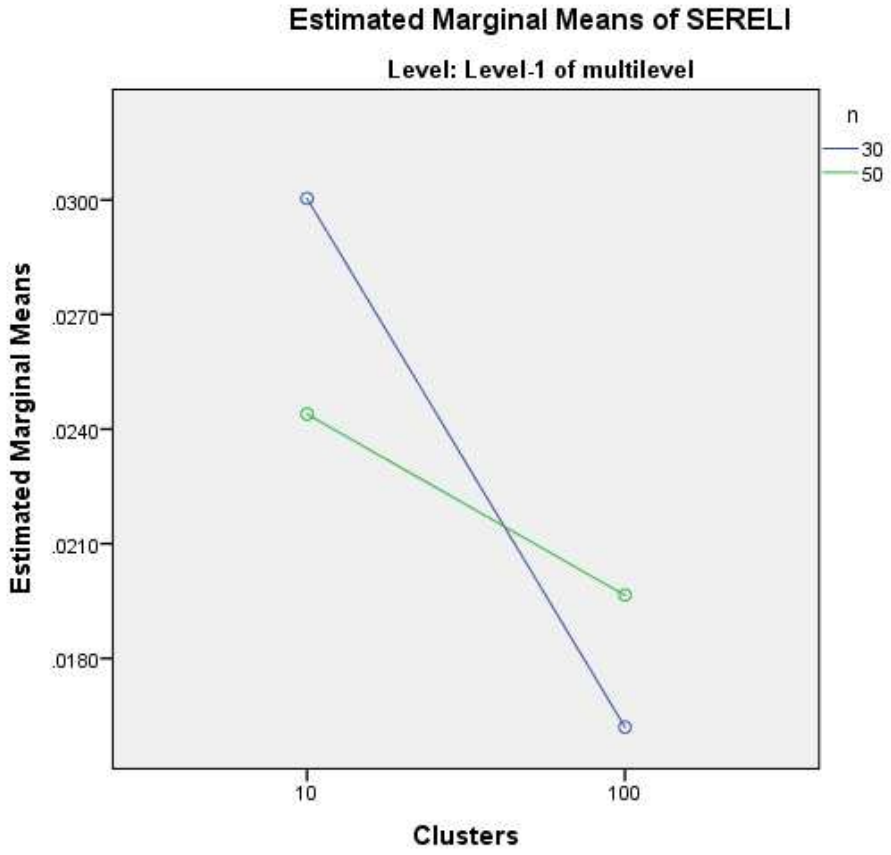

Figure 19. The interaction plot showing the interaction between level-1 and level-2 sample sizes on estimates of standard errors of reliability (SERELI).

Table 29 shows the results of the simple main effects when the interaction plots do not show any discernible interactions even though they were found to be statistically significant. Based on the effect sizes for two way interactions reported in Table 29, the results suggested that no substantial differences in level-1 standard errors existed that depensws on interactions between independent variables.

Interactions for Standard Errors at Level-2

With α =.05, all interactions were statistically significant. The plot of interactions presented no discernible interactions for (a) Level-1 * Level-2 sample sizes, (b) Level-1 sample size * Type of reliability and (c) Type of reliability * Distribution and the effect sizes for each was > .02; therefore, following the same procedures of analysis described above when the dependent variable was level-1 standard errors, I report the significant interactions even if the effects were negligible.

Figure 20 shows the effects on level-2 standard errors based on the interaction between level-2 sample sizes and type of reliability, with higher level-2 reliability estimates for both Cronbach's α ($F[1, 35971] = 544.641$, $p < .0001$, $\eta^2 = .014$), and person reliability when level-2 sample size $N = 10$ (i.e., number of clusters is 10), ($F[1, 35971] = 128.168$, $p < .0001$, $\eta^2 = .004$) but higher polychoric ordinal α when the number of clusters $N = 100$.

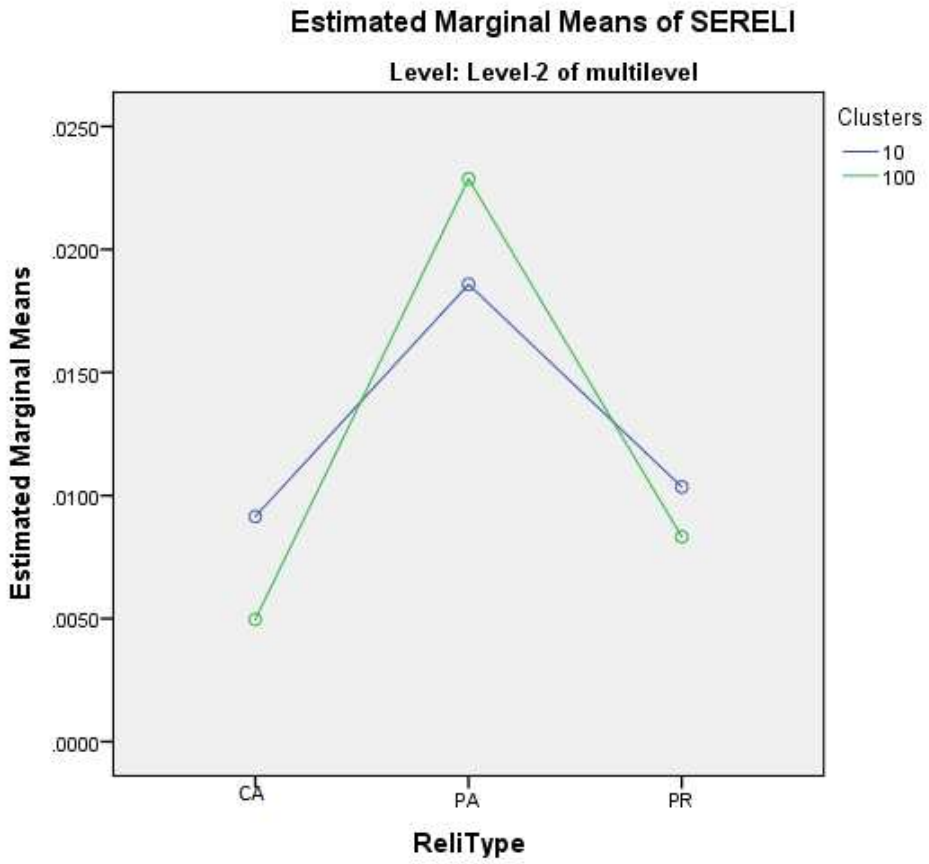

Figure 20. Plot of the interaction effects between level-2 sample size and type of reliability (ReliType) for level-2 standard errors.

Level-1 Percentage of Relative Bias Across Data Conditions

An examination of the percentage of level-1 relative bias ≥ 10% across all data conditions yielded the results presented in Table 30 below. Only two two-way interactions were statistically significant (type of reliability * level-2 sample size and level-1 sample size * level-2 sample size) and neither had an effect size > .01; therefore, I did not interpret the interactions. Only two main effects showed statistical significance: Type of Reliability and Level-2 sample size. Type of reliability results (F[2, 35972] =

1391.822, $p < .0001$, $\eta^2 = .068$) show that person reliability at level-1 of a two-level model was significantly higher than Cronbach's α and polychoric ordinal α based on a Bonferroni post hoc comparison. Type of reliability has a moderate effect on the amount of bias found in level-1 reliability coefficients while level-2 sample size has no substantial effect.

Level-2 Percentage of Relative Bias Across Data Conditions

A factorial ANOVA was conducted with level-2 bias as the dependent variable and type of reliability, data distribution, and level-1 and level-2 sample sizes as the independent variables. Since, as you may note from Table 23, reported, where Spearman-Brown reliability coefficient bias is, no discernable bias existed in level-2 reliability coefficients across all data conditions.

Direction of Relative Bias Across Data Conditions

Table 31 shows the actual number of biased reliability estimates across all data conditions (out of 1,000 for each condition). One unexpected result presented in Table 31 that stood out was related to polychoric ordinal α under the condition of a mixed distribution. For level-2 sample size $N = 100$, and level-1 sample size $N = 30$, 583 out of 1,000 reliability coefficients did not only demonstrate bias $\geq 10\%$, but were underestimated; however, when level-1 sample size $N = 50$, 453 out of 1,000 reliability coefficients demonstrating bias $\geq 10\%$ were overestimated. This irregularity is discussed further in Chapter V.

Table 32 shows the Chi-square test results for the direction of relative bias $\geq 10\%$ found in level-1 across all data conditions. In other words, if a set of reliability

coefficients (based on data conditions) had bias ≥ 10%, then whether the bias was overestimated or underestimated was determined. Direction of the bias was the dependent (categorical) variable and type of level-1 reliability coefficient, level-1 and level-2 sample sizes, and data distributions were the other variables in the Chi-square tests. There were 11,772 reliability coefficients (out of 36,000) with relative bias ≥ 10% and 6,857 were overestimated and 4,915 were underestimated.

The results supported my hypotheses that in multilevel models, bias in reliability estimates in level-1 would increase under the conditions of smaller level-1 and level-2 sample sizes and mixed and non-normal distributions. My hypothesis stating that polychoric ordinal α would be less biased than Cronbach's α and person reliability was only partially supported. Polychoric ordinal α was less biased under non-normal data conditions, however; showed a higher bias ≥ 10% (average bias = 30.74%) than Cronbach's α bias ≥ 10% (average bias = 20.39 while person reliability bias ≥ 10% was greater than both across data conditions. Finally, my hypothesis regarding Spearman-Brown coefficients being underestimated under the conditions of smaller sample size and non-normal or mixed distributions was not supported, with the exception of high relative bias found under the condition of a mixed distribution, level-1 sample size $N = 30$ and level-2 sample size $N = 100$. Otherwise, Spearman-Brown coefficients showed little to no bias across data conditions. The anomaly of the high relative bias mentioned above is discussed further in Chapter V.

A Comparison of Single-Level and Level-1 Standard Errors and Bias Across Data Conditions

Q3 Do standard errors and percentage of bias ≥ 10% differ between types of reliability estimates at the single-level and at level-1 of a two-level model regardless of sample size and distribution of data (a comparison of Cronbach's α, polychoric ordinal α, and person reliability)?

H3 Cronbach's α, polychoric ordinal α, and person reliability standard errors and percentage of bias ≥ 10% for level-1 of the two-level model will be lower than the standard errors and bias found in the single-level models regardless of the sample size or distribution of data.

To provide a reasonable comparison of standard errors and percentage of bias in single-level and level-1 of two-level sampling designs across data conditions, the sample size used in the single-level model was $N = 300$ and in the two-level model, I used 10 clusters of 30 for a total of 300 clusters.

Standard errors. Table 33 shows the average standard errors across types of reliability, sample sizes, and data distributions in both the single-level and level-1 of a two-level model. Table 33 shows that the standard errors in a single-level model range between .0351 and .1822 and in level-1 of a two-level model range between .0275 and .0406. As explained by the central limit theorem and evidenced in these data, as sample size increases, standard errors decrease. In addition, the magnitude of standard errors depend upon the distribution of data as well as the sampling design.

Table 33

Average Standard Errors of Measurement for Reliability Estimates Across Data Conditions

Sampling Design	Type of Reliability Coefficient	Distribution	Sample Size	Average SE
Single-Level	PA	Mixed	30	.0576
	PR	Non-Normal	30	.1882
	CA	Normal	50	.0558
	PA	Mixed	50	.0519
	PR	Non-Normal	50	.0882
	CA	Normal	300	.0351
	PA	Mixed	300	.0802
	PR	Non-Normal	300	.0830
Level-1	CA	Normal	300*	.0291
	PA	Mixed	300*	.0406
	PR	Non-Normal	300*	.0275

* represents 30 individuals in 10 groups for comparison to single-level model.

Reliability bias. Table 34 shows the amount of and percentage of average absolute relative bias ≥ 10% in the single-level and level-1 of the two-level sampling design across the data conditions. Two key findings stand out: (a) The amount of average relative bias in single-level models is higher across most data conditions, with the exception of Cronbach's α under the condition of a normal distribution and $N = 300$ and person reliability under the condition of a mixed data distribution and $N = 300$. (b) The percentage of average relative bias ≥ 10% is higher for single-level sampling designs than

for multilevel sampling design when compared across most data conditions with the exception of Cronbach's α under the condition of a normal distribution and $N = 300$ and person reliability under the condition of a mixed data distribution and $N = 300$.

Test of Hypotheses for Research Question 3

Two factorial ANOVAs and one Chi-Square test were conducted to answer research question three. For the factorial ANOVAs, the dependent variables were *standard errors* and *percentage of bias ≥ 10%.* The independent variables were the sampling design (level: single-level and level-1) type of reliability coefficient (CA, PA, and PR), sample size (single-level $N = 30, 50, 300$, level-1 $N = 300$), and type of data distribution (normal, mixed, and non-normal). A Chi-square test was conducted to analyze the direction of absolute relative bias since the dependent variable was categorical (overestimated or underestimated). Following a review of the statistical and graphical output generated for the factorial ANOVAs, I found no statistically significant interaction effects relative to the dependent variable of standard errors or percentages of bias; therefore main effects results, presented in Tables 35 and 36, were examined. All results were statistically significant, with the exception of bias based on sample size.

Table 34

A Comparison of Single-Level Bias to Level-1 Bias Across Data Conditions

	Single-Level	Level-1	Single-Level	Level-1
Sample Size	Average Reliability Coefficient Relative Bias	Average Reliability Coefficient Relative Bias	Percentage of Bias ≥ 10%	Percentage of Bias ≥ 10%
Cronbach's α				
All Data Normally Distributed				
30	.104	N/A	39.50%	N/A
50	.076	N/A	26.90%	N/A
300	.029	.042	0.50%	4.20%
Mixed Data Distribution				
30	.159	N/A	78.20%	N/A
50	.172	N/A	99.80%	N/A
300	.225	.043	98.50%	24.30%

Table 34 (continued)

Sample Size	Single-Level Average Reliability Coefficient Relative Bias	Level-1 Average Reliability Coefficient Relative Bias	Single-Level Percentage of Bias ≥ 10%	Level-1 Percentage of Bias ≥ 10%
Non-Normal Distribution				
30	.175	N/A	55.90%	N/A
50	.149	N/A	56.00%	N/A
300	.173	.042	94.40%	21.60%
Polychoric Ordinal α				
All Data Normally Distributed				
30	.113	N/A	41.20%	N/A
50	.082	N/A	29.90%	N/A
300	.032	.026	2.00%	27.60%

Table 34 (continued)

Sample Size	Single-Level Average Reliability Coefficient Relative Bias	Level-1 Average Reliability Coefficient Relative Bias	Single-Level Percentage of Bias ≥ 10%	Level-1 Percentage of Bias ≥ 10%
Mixed Data Distribution				
30	.162	N/A	79.00%	N/A
50	.172	N/A	68.60%	N/A
300	.136	.053	99.70%	35.90%
Non-Normal Distribution				
30	.475	N/A	96.60%	N/A
50	.129	N/A	62.40%	N/A
300	.111	.042	55.90%	14.00%

Table 34 (continued)

Sample Size	Single-Level Average Reliability Coefficient Relative Bias	Level-1 Average Reliability Coefficient Relative Bias	Single-Level Percentage of Bias ≥ 10%	Level-1 Percentage of Bias ≥ 10%
Person Reliability (RSM)				
All Data Normally Distributed				
30	.163	N/A	84.00%	N/A
50	.208	N/A	100.00%	N/A
300	.212	.217	100.00%	99.80%
Mixed Data Distribution				
30	.157	N/A	18.00%	N/A
50	.029	N/A	2.00%	N/A
300	.000	.056	0.00%	99.90%

Table 34 (continued)

Sample Size	Single-Level Average Reliability Coefficient Relative Bias	Level-1 Average Reliability Coefficient Relative Bias	Single-Level Percentage of Bias ≥ 10%	Level-1 Percentage of Bias ≥ 10%
Non-Normal Distribution				
30	.386	N/A	100.00%	N/A
50	.414	N/A	100.00%	N/A
300	.457	.031	100.00%	17.90%

Table 35

Single-Level and Level-1 Standard Errors of Reliability Estimates Across Data Condition: Assessment of Main Effects

Source	*F**	Post-hoc Analysis/Explanation	Eta squared
Distribution	F = 30340.437	non-normal distributions have higher standard errors than normal or mixed distributions	.047
Level (single-level or level-1)	*F = 16165.073*	single-level standard errors are higher than standard errors in level-1 of a twp-level model	.310
Type of Reliability	*F = 27103.855*	standard errors are lower for person reliability than for Cronbach's a or polychoric ordinal a	.041
Sample Size	F = 1273.85	standard errors were lower for N =300 than for N=30 or N = 50	.002

*p < .0001

Table 36

Results of Factorial ANOVA Comparing Bias ≥ 10% Between Single-Level and Level-1 of a Two-Level Sampling Design

IV's	Levels	*F* Statistic	eta squared
Single or Level-1	Single-Level or Level-1 of a two-level model	45.596	.007
Distribution	Normal, Mixed, & Non-Normal	64.422	.210
Sample Size	*N* = 30, 50, 300	3.32	.009
Type of Reliability	Cronbach's a, polychoric ordinal a, and person reliability	9.739	.030

Note. $p < .0001$

Although the main effects results were statistically significant, effect sizes were small for distribution, type of reliability, and sample size, indicating the differences in standard errors across these data conditions was not substantial. There were statistical and substantial differences in standard errors between the single-level and level-1 of a two-level sampling design, with lower standard errors found at level-1 of a two-level sampling design.

Table 36 shows the results of the factorial ANOVA assessing bias between single-level and level-1 of two-level models. Main effects are reported since no statistically significant interactions were present. While all data conditions were statistically significant at the $\alpha = .05$ level, only data distribution has an effect size > .031, which indicates distribution has a small effect on the percentage of bias in the single-level and level-1 of a two-level model. The effect sizes calculated indicate that the amount of variance in bias is not well explained by the sampling design.

Recall that in both single-level and level-1 of a two-level model, reliability was fixed at .70 for Cronbach's α, polychoric ordinal α, and person reliability, allowing a comparison of these sampling designs. If the generated sample reliability coefficient was > the known reliability coefficient of .70, then it was considered overestimated. Conversely, if the generated sample reliability coefficient was < the known reliability coefficient of .70, then it was considered underestimated. To assess whether reliability estimates with average absolute relative bias ≥ 10% were underestimated or overestimated, a Chi-square test with the categorical variable *level* as the dependent was conducted and the results are in Table 37 below.

Table 37

Single-Level and Level-1 Direction of Relative Bias ≥ 10% Across Distribution and Sample Sizes (Chi-square)

Source	*df*	Chi-square	*p*-value
Distribution	4	1620.06	$p < .0001$
Type of Reliability	6	18288.50	$p < .0001$
Level	1	5002.36	$p < .0001$
Sample Size	3	2.847	$p < .20$

The direction of average absolute relative bias ≥ 10% differed significantly between single-level and level-1 sampling designs. Of reliability estimates with average absolute relative bias ≥ 10%, overestimation of these reliability estimates occurred 63.53% of the time in single-level models and 36.47% of the time in level-1 of a two-level model. Of reliability estimates with average absolute relative bias ≥ 10%, underestimation of these reliability estimates occurred 23.26% of the time in single-level models and 76.74% of the time in level-1 of a two-level model. Cronbach's α and polychoric ordinal α demonstrated the tendency to underestimate reliability coefficients while person reliability overestimated reliability estimates.

The data provide evidence to support my hypothesis that standard errors of measurement in Cronbach's α, polychoric ordinal α, and person reliability are lower for level-1 of the two-level model than for single-level models. My hypothesis that average absolute relative bias ≥ 10% would be smaller in level-1 of a two-level model when compared to a single-level model was supported. While statistically significant, effect sizes across data conditions were small, indicating the results either make very little

difference on reliability estimates or they need a theoretical framework to provide context.

Single-Level and Multilevel Interactions Across Data Conditions

Q4 To what degree do interactions among sample size, data distribution, and sampling design (e.g., single-level and two-level) affect levels of bias in reliability estimates (a comparison of bias in Cronbach's α, polychoric ordinal α, person reliability, and Spearman-Brown coefficients)?

H4 Two-way Interactions among sample size, data distribution, and sampling design will increase bias in reliability estimates, with the joint effects of lower sample sizes and non-normal and/or mixed distributions displaying the most bias.

Across both single and two-level models, the distribution of data had the largest effect on the percentage of average absolute relative bias ≥ 10% across data conditions; however, in the single-level model, no statistically significant interaction effects were found. This means that in single-level models, the effect of one independent variable did not depend on the value or level of another independent variable included in the model (Jaccard & Turrisi, 2003; Oshima & MCcarty, 2015). Numerous interaction effects were found in the multilevel model as described below as detailed above for Research Questions 2 and 3. This means that the effect of one independent variable was dependent on the value or level of some other independent variable included in the study design (Jaccard & Turrisi, 2003; Oshima & MCcarty, 2015). For example, as seen in Table 25 above, the effect of level-1 sample size on level-1 reliability coefficients depends on the type of data distribution. In Table 27 presented above, the effect of level-1 sample size on level-2 reliability coefficients depend upon a combination of distribution and type of reliability. As presented in Table 28, presented previously and mentioned here as a comparison, complex level-1 interactions existed between data conditions. Untangling

and interpreting these complex interaction effects showed that the effect of level-1 and level-2 sample sizes was dependent on the data distribution and type of reliability coefficient. Normal and mixed distributions, as well as person reliability, showed more bias than non-normal distributions. These results were driven by the fact person reliability bias in non-normal distributions were unexpectedly low, due to such large standard errors. This is examined further in Chapter V. Finally, no interaction effects were present regarding level-2 bias. In addition, level-2 (Spearman-Brown) reliability estimates were low and stable with the exception of high relative bias under the conditions of mixed data distributions at level-1 $N = 30$.

My hypothesis that interactions among sample size, data distribution, and sampling design would increase bias in reliability estimates was not supported in single-level models as no interaction effects existed. This hypothesis was partially supported in multilevel models since the effects of level-1 and level-2 sample size was dependent on the data distribution and type of reliability coefficient, and normal and mixed distributions, as well as person reliability, showed more bias than non-normal distributions.

Conclusions

While most single-level and multilevel estimates of reliability and percentage of the absolute value of relative bias ≥ 10% were expected, several single-level and multilevel results regarding bias were unexpected and are explored further in Chapter V.

Single-Level Sampling Designs

Cronbach's α and polychoric ordinal α reliability coefficients, along with their associated standard errors of measurement provided similar results across data conditions

while person reliability coefficients were high and standard errors were low under conditions of non-normality. Reliability coefficients were fairly stable across mixed data distributions and standard errors were low and ranged between .02 and .06, and reliability coefficients were low across non-normal data distributions and standard errors were high as seen in Figure 21. For example, in Figure 21, presented below, average single-level reliability coefficients are on the Y-axis, average standard errors are on the X-axis, and the data distribution and type of reliability coefficient are shown with results by sample size. For example, the graph shows that under the conditions of a normal distribution, reliability coefficients range between .60 and .90 and standard errors are low across type of reliability and sample sizes. Reading the graph from top to bottom and left to right, the bottom right cell shows that person reliability estimates were low while standard errors were unusually high.

Normal distributions. Bias and direction of bias in Cronbach's α and polychoric ordinal α were similar under normally distributed data conditions where bias decreased as sample size increased and reliability estimates were underestimated significantly more often than they were overestimated. Person reliability under normal data distributions showed high bias and was significantly overestimated.

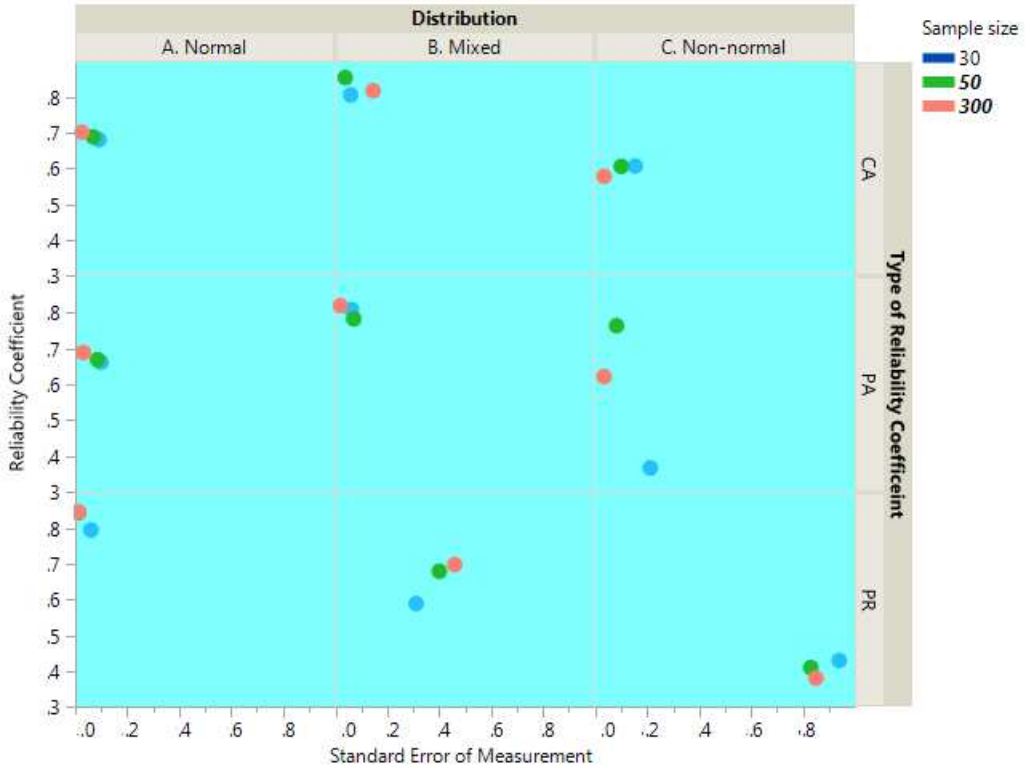

Figure 21. Cronbach's α and polychoric ordinal α reliability coefficients, along with their associated standard errors of measurement across data conditions.

Mixed distributions. There are some discrepancies in bias in Cronbach's α and polychoric ordinal α under mixed data distributions, where bias increased as sample size increased for Cronbach's α and decreased at $N = 300$ for polychoric ordinal α, and where both were overestimated significantly more often than underestimated. Under mixed data distributions, bias decreased as sample size increased and at $N = 300$ no bias was visible. Where bias was seen, it was more often overestimated than underestimated. These results indicate that polychoric ordinal α provides a less biased estimate of reliability than Cronbach's α under mixed distributions and larger sample sizes and person reliability provides the most unbiased estimates of reliability under these data conditions.

Non-normal distributions. Under non-normal data conditions, bias in Cronbach's α and polychoric α is high and significantly more often underestimated than overestimated. Two noteworthy results are that polychoric ordinal α had unusually high bias when $N = 30$ and severely overestimated bias when $N = 50$. These results are explored further in Chapter V. Bias in person reliability was extremely high and underestimated 100% of the time.

Conclusions for single-level results. A broad view of the single-level results indicates that Cronbach's α and polychoric α provide good estimates of reliability across sample sizes and under the condition of normally distributed data. Under conditions of mixed and non-normal data distributions, person reliability provides the best estimates of reliability across sample sizes.

Multilevel Sampling Designs

Complex interaction effects are found in Level-1 of multilevel models, making interpretation difficult. After reviewing interaction plots and assessing simple main effects, the emerging results showed that Cronbach's α and polychoric ordinal α provided (a) similar reliability estimates and standard errors across data distributions as well as across cluster and cluster sample sizes; however, person reliability overestimated reliability under conditions of normally distributed data across these same conditions. (b) person reliability was far more accurate under conditions of mixed and non-normal distributions across cluster and cluster sample sizes. (c) the standard errors of person reliability estimates were similar to those found in Cronbach's α and polychoric ordinal α across all data conditions.

Normal distributions. Bias in Cronbach's α and polychoric ordinal α at level-1 was similar under normally distributed data conditions where bias remained low and stable across cluster and cluster sample sizes; however, bias in Cronbach's α was significantly more often underestimated while bias in polychoric ordinal α was more often overestimated. Bias and the direction of bias in person reliability estimates was similar to polychoric ordinal α when the number of clusters was 100 and higher when the number of clusters was only 10 and cluster sample size was only 30.

Mixed distributions. Bias in Cronbach's α and polychoric ordinal α under mixed data distributions showed similar results where it was low and underestimated. Person reliability bias decreased as cluster sample size increased when the number of clusters was10 and conversely, bias in person reliability increased as sample size increased when the number of clusters was 100.

Non-normal distributions. Under non-normal data conditions, bias in Cronbach's α and polychoric α is higher when the number of clusters was higher (N = 100) and significantly more often underestimated than overestimated. Bias in person reliability was lower when the number of clusters was 100 compared to when there were only 10 clusters and significantly lower than both Cronbach's α and polychoric ordinal α

Conclusions for multilevel results. A broad view of the multilevel results in level-1 of two-level models indicates that Cronbach's α and polychoric α provide good estimates of reliability across sample sizes and under the condition of normally distributed data. Under conditions of mixed data distributions, polychoric ordinal α provides the least biased reliability coefficients and under non-normal data distributions, person reliability provides the best estimates of reliability across sample sizes. Spearman-

Brown provides stable and accurate estimates of between level reliability coefficients with extremely low bias.

Overall, reliability estimates, standard errors and bias improve in multilevel sampling designs when compared to single-level sampling designs across normal distributions, types of reliability, and cluster and cluster level sample sizes. These results support the results from B. O. Muthén (1994), B. Muthén and Asparouhov (2011), Raykov and Penev (2010), and Snijders and Bosker (1999). Further, these results also support recommendations by Kamata (2001) and Raudenbush et al. (2003) about using multilevel sampling designs in IRT models. Further, results support Snijders and Bosker (1999) who suggested that if single-level models are used to assess parameters when multilevel sampling designs were employed, within-cluster variance and between-cluster variance would be confounded, leading to relatively biased reliability estimates in single-level models since the assumption of independent residuals is violated. Finally, in the multilevel CTT framework, Cronbach's α was slightly less biased than polychoric ordinal α at level-1 of the two-level model under normal data distributions while polychoric ordinal α was less biased than Cronbach's α under mixed and non-normal data distributions. The multilevel Rasch modeling framework showed significantly less bias in person reliability estimates at level-1 of a two-level model under mixed and non-normal data distributions than either Cronbach's α and polychoric α. Several unexpected data anomalies were seen during the analyses of these data which are explored in more detail in Chapter V.

CHAPTER V

CONCLUSIONS

The principle focus of this dissertation was to assess the amount of bias in Cronbach's α, polychoric ordinal α, and person reliability estimates found in polytomously scored items in a multilevel sampling design under mixed and non-normal data distributions. Gadermann et al. (2012) and Geldhof et al. (2014) advised examining polychoric ordinal α under varying sample sizes and distributional characteristics in multilevel models, Raykov and Penev (2010) and Sheng and Sheng (2012) suggested measuring corresponding levels of reliability bias under conditions of multilevel sampling designs and non-normal distributions, and Huang and Cornell (2016) proposed assessing reliability coefficients derived from polytomously scored items under non-normal distributions.

To meet these challenges, I used Monte Carlo simulation techniques to generate multivariate data representing normal, mixed, and non-normal distributions, varying individual and cluster level sample sizes, and single- and two-level sampling designs. Simulating the polytomously scored observed responses for this study allowed me to control the data characteristics and concentrate my analysis on the specific data conditions recommended by Gadermann et al. (2012), Geldhof et al. (2014), Raykov and Penev (2010), Sheng and Sheng (2012), and Huang and Cornell (2016).

I generated multivariate data for both single- and two-level sampling designs to compare the behavior of and measure bias across three reliability estimates with the

primary objective of offering basic and applied researchers, clinicians, and educators recommendations on the least biased reliability coefficients in both single and multilevel sampling designs.

To meet my goal of measuring the amount of bias found in Cronbach's α, polychoric ordinal α, and person reliability estimates, I computed the reliability estimates across all data conditions, calculated the corresponding standard errors and 95% confidence intervals for these estimates, and, based on the advice of B. Muthén and Kaplan (1985) and Geldhof et al. (2014), calculated and recorded relative bias ≥ 10%. Through these processes, I encountered both expected and unexpected results in single and multilevel models and I highlight the key findings by sampling design below.

Research Question 1: Single-Level
Results and Discussion

Expected Results

Cronbach's α and polychoric ordinal α provided similar estimates of reliability and standard errors under normal and mixed data distributions. Sample estimates of Cronbach's α and polychoric ordinal α increased and drew closer to the known fixed reliability of .70, and standard errors decreased, as sample size increased. Cronbach's α and polychoric ordinal α showed similar levels of relative bias under normal and mixed data distributions, which decreased as sample sizes increased, and both coefficients underestimated reliability under the condition of normally distributed data.

Unexpected Results

An unusual pattern emerged in mixed data distributions for Cronbach's α, polychoric ordinal α, and person reliability across sample sizes. Cronbach's α and

polychoric ordinal α overestimated reliability coefficients and person reliability underestimated reliability coefficients under these data conditions. Boomsa (1983) found that skew > 2 tended to overestimate Cronbach's α reliability estimates while Linacre (2014) explained that reliability estimates that include extreme scores, as is the case in the mixed distributions, are usually lower than when scores meet the definition of normality. Furthermore, standard errors of reliability estimates for Cronbach's α and polychoric ordinal α ranged between .06 and .144 and for person reliability, ranged between .31 and .46, which indicates person reliability estimates could not capture the "true" score as efficiently as Cronbach's α or polychoric ordinal α in a mixed distribution. Another unexpected result was that person reliability was overestimated under conditions of normally distributed data across sample sizes with high standard errors for mixed and non-normal distributions. Linacre (2014) stated that person reliability, when compared to Cronbach's α, tended to be underestimated in the Rasch IRT model. I found the opposite to be true in the data simulated for this dissertation. I reviewed the data generation techniques for any anomalies and found none. I then reviewed the literature and found Linacre (2017) suggested that the higher number of response categories would translate into higher person reliability estimates for smaller sample sizes and normally distributed data. Further analysis revealed that Zhang (2010) examined the issue of overestimated reliability estimates in the Rasch rating scale model (RSM) and explained that person reliability was overestimated at smaller sample sizes, which Zhang considered to be under 500. This may explain the overestimation of person reliability under the conditions of normally distributed data and sample sizes of 30, 50, and 300.

To explain the overestimation of standard errors in mixed and non-normal distributions, I turned to Daher, Ahmad, Winn, and Selamat's (2015) study regarding the effect of standard errors on reliability estimates in Rasch rating scale models. The authors explained that the standard errors used in the calculation of CTT reliability coefficients are derived from the average of sample ability while in the RSM model, the person reliability coefficient is based on individual person ability. Therefore, the reliability error variance in the RSM model may be overestimated, especially under conditions of mixed and non-normal distributions, which is supported by my results.

Research Questions 2 through 4: Multilevel Results and Discussion

Expected Results

Assessing reliability estimates derived from the multilevel model, polychoric ordinal α provided slightly more precise estimates of reliability across all data conditions than Cronbach's α and significantly more precise estimates than person reliability under normal and mixed data distributions in level-1 of a two-level model. Furthermore, the corresponding standard errors in level-1 of a two-level model were substantially lower than the standard errors of a single-level model. Finally, significantly less bias was found in level-1 reliability estimates than in reliability estimates derived from single-level models.

Two concepts may intersect to explain these results. The first is based on Raudenbush and Bryk's (1994) study where they found that in multilevel models, level-1 standard errors had *downward bias.* Second, Maas and Hox (2005) found that smaller level-1 standard errors are due to the partitioning of error variance at the within and between levels of a two-level model. In other words, single-level models have more error

due to the fact clustering effects are not taken into account, whereas multilevel models spread the error across more than one level. It is noteworthy to add a caveat that may affect interpretation of results regarding parameter estimates. Raudenbush and Bryk (1994) rightly pointed out that smaller standard errors in multilevel models may lead to a Type I error. They explained that standard errors can be too small to provide interpretable results, especially when assessing reliability, which is measured between 0 and 1.

Unexpected Results

Relative bias for person reliability under the conditions of normally distributed data, level-1 samples sizes N = 30 and level-2 sample size N =10 (i.e., number of clusters) was unexpectedly high (Relative bias = .217). This result was an artifact of the high average reliability estimate found in level-1 of the two-level model, which is partially explained by Zhang (2010) who attributed high person reliability estimates in RSM models to small sample sizes; however, this does not explain why the average relative bias was < 6% across all other person reliability data conditions. Future researchers may want to examine this anomaly in more detail.

Spearman-Brown coefficients had almost no bias, meaning that level-2 (between-level) reliability estimates were decidedly accurate across types of level-1 reliability coefficients, level-1 and level-2 sample sizes, and data distributions. The Spearman-Brown prophecy formula represents the proportion of level-1 scores accounted for by level-2 membership. Therefore, more precise estimates of Spearman-Brown coefficients demonstrate the importance of a multilevel model in the analysis of data.

Less average relative bias was seen in person reliability under non-normal data distributions than found in Cronbach's α or polychoric ordinal α under the same data

conditions. This result was an artifact of the more precise average reliability estimates computed for person reliability under conditions of non-normality. Although the skew was > 2 in the non-normal distribution, the total level-1 and level-2 sample sizes likely made up for the underling distribution of data. In a two-level model, sample sizes were between 300 to 5,000 individuals across clusters, which Zhang (2010) demonstrated would provide more precise estimates of reliability. The data generated for this dissertation support Zhang's conclusions.

Finally, inconsistent results for polychoric ordinal α under the condition of a mixed distribution were revealed in the data generated for this dissertation. For level-2 sample size N = 100, and level-1 sample size N = 30, 583 out of 1,000 reliability coefficients demonstrating bias ≥ 10% were underestimated; however, when level-1 sample size N = 50, 453 out of 1,000 reliability coefficients demonstrating bias ≥ 10% were overestimated. The reasons are not clear and may just be an artifact of the characteristics of the mixed distribution. For example, since one fundamental assumption of these data was unidimensionality, the mixed distribution of data was accomplished by generating the first five items using a polychoric correlation matrix and multivariate distribution in R. The next five items were generated using a polychoic correlation matrix and an extremely non-normal distribution in R, with skew =3 and kurtosis = 7. Had mixed data been generated as 50% of the respondents representing a normal distribution, and 50% of the respondents representing an extreme non-normal distribution, the assumption of unidimensionality would have been violated. These types of data generation methods may have contributed to anomalies in the mixed distributions which were more apparent for polychoric ordinal α than for Cronbach's α or person reliability.

Implications and Recommendations

During the development phase of any assessment tool, reliability and validity are considered “the two most important fundamental characteristics of any [psychometric] procedure” (Miller, 2004, p. 1). Miller (2004) explained that scores on an assessment instrument can be reliable (representing consistency and reproducibility) without being valid (representing accuracy) but cannot be valid without first being reliable. Therefore, understanding the amount of relative bias found in Cronbach’s α, polychoric ordinal α and person reliability estimates across data conditions is essential. Although most behavioral, educational, and social science data have a hierarchical structure (e.g., students nested within schools or patients nested within clinics), most researchers ignore the clustered nature of the data and use single-level modeling techniques to assess their results which suggests more research on these topics needs to be conducted.. Therefore, I utilized Monte Carlo simulation techniques in this dissertation to provide researchers, educators, and clinicians with more clarity regarding the computation and interpretation of reliability estimates derived from the scores on an assessment instrument, survey, or questionnaire.

The results of my dissertation support the recommendation of taking the structure of the data collected into account during the analytic phase made by T. A. Brown (2015), Gadermann et al. (2012), Geldhof et al. (2014), Huang and Cornell (2016), Nunnally and Bernstein (1994), Raudenbush and Bryk (2002), Raykov and Penev (2010), and Sheng and Sheng (2012). The results reported in this dissertation provide empirical evidence that if data collected for research are dependent on a higher order structure (such as students nested within schools), reliability coefficients in a multilevel model are less

biased than reliability coefficients derived from a single-level model. Additionally, results support the idea that polychoric ordinal α at level-1 of a two-level sampling design provided slightly more precise estimates of reliability across all data conditions than Cronbach's α and significantly more precise estimates than person reliability under normal and mixed data distributions; however, the small gain in the precision of reliability estimates may not be worth the additional effort of using polychoric correlation matrices to estimate reliability for many clinicians and educators. Consequently, using Cronbach's α under normal and mixed data conditions and across sample sizes is certainly acceptable, and far easier to estimate since it is available in most statistical software packages used in the social sciences. If behavioral, educational, and social science researchers and applied practitioners find their data to be extremely non-normal, my recommendation is to estimate reliability using the Rasch-RSM model since the effort to estimate reliability using the Rasch-RSM is worth the lower level of bias found under these conditions and across sample sizes. Finally, if computing either polychoric ordinal α or person reliability using the Rasch-RSM model causes extreme distress, Cronbach's α is a good alternative under normal or non-normal distributions in both single and multilevel sampling designs as long as it is understood that Cronbach's α is likely to be underestimated across data conditions and the results are reported inappropriately. Cronbach's α is not a good choice for mixed data distributions in a multilevel model and should be avoided. I propose calculating and reporting polychoric ordinal α. Tables 38 and 39 represent a tool that social science researchers can use to determine the most appropriate reliability coefficient to report based on level-1 and level-2 sample size and type of distribution, as well as the consideration of effort vs. benefit. The results also

show that a variety of different data properties, including data distribution and sample size, jointly affect reliability coefficients and care should be taken, not only to provide context to the data structure, but also a theoretical framework in which to interpret the results. Tables 38 and 39 are tools developed to guide applied social science researchers in their decisions regarding which reliability coefficient to report and how to compute that coefficient. The tools are based upon the most expeditious coefficient to calculate under each set of data conditions, taking into account any differences in levels of bias versus the effort of the computation and explanation. Table 38 represents recommendations for a single-level model and Table 39 represents recommendations for a two-level model.

Table 38

A Single-Level Model Tool for Applied Researchers

Sample Size	Type of Distribution	Reliability Coefficient Recommendations	Measurement Framework	Recommended Statistical Software
30	Normal	Cronbach's α	Classical Test Theory	SPSS, SAS, STAT, R
	Mixed			
	Non-Normal (skew and kurtosis = +/- 2)			
	Extremely Non-Normal (skew = +/- 3, kurtosis = +/- 7)	Person Reliability	Item Response Theory (Rating Scale Model)	Winsteps, R
50	Normal	Cronbach's α	Classical Test Theory	SPSS, SAS, STAT, R
	Mixed			
	Non-Normal (skew and kurtosis = +/- 2)			
	Extremely Non-Normal (skew = +/- 3, kurtosis = +/- 7)	Person Reliability	Item Response Theory (Rating Scale Model)	Winsteps, R

Table 38 (continued)

Sample Size	Type of Distribution	Reliability Coefficient Recommendations	Measurement Framework	Recommended Statistical Software
300	Normal	Cronbach's α	Classical Test Theory	SPSS, SAS, STAT, R
	Mixed			
	Non-Normal (skew and kurtosis = +/- 2)			
	Extremely Non-Normal (skew = +/- 3, kurtosis = +/- 7)	Person Reliability	Item Response Theory (Rating Scale Model)	Winsteps, R

Table 39

A Two-Level Model Tool for Applied Researchers

Size Level 1	Size Level 2	Distribution	Coefficient Recommendations (level-1 Within)	Framework	Statistical Software
30	10	Normal	Cronbach's α	Classical Test Theory	SPSS, SAS, R, STATA
		Mixed			
		Non-Normal (skew and kurtosis = +/- 2)			
		Extremely Non-Normal (skew = +/- 3, kurtosis = +/- 7)	Person Reliability	Item Response Theory (Rating Scale Model)	Winsteps, R
50	10	Normal	Cronbach's α	Classical Test Theory	SPSS, SAS, R, STATA
		Mixed	Polychoric ordinal α		
		Non-Normal (skew and kurtosis = +/- 2)	Cronbach's α		
		Extremely Non-Normal (skew = +/- 3, kurtosis = +/- 7)	Person Reliability	Item Response Theory (Rating Scale Model)	Winsteps, R

Table 39 (continued)

Size Level 1	Size Level 2	Distribution	Coefficient Recommendations (level-1 Within)	Framework	Statistical Software
30	100	Normal	Cronbach's α	Classical Test Theory	SPSS, SAS, R, STATA
		Mixed	Polychoric ordinal α		
		Non-Normal (skew and kurtosis = +/- 2)	Cronbach's α		
		Extremely Non-Normal (skew = +/- 3, kurtosis = +/- 7)	Person Reliability	Item Response Theory (Rating Scale Model)	Winsteps, R
50	100	Normal	Cronbach's α	Classical Test Theory	SPSS, SAS, R, STATA
		Mixed	Polychoric ordinal α		
		Non-Normal (skew and kurtosis = +/- 2)	Cronbach's α		
		Extremely Non-Normal (skew = +/- 3, kurtosis = +/- 7)	Person Reliability	Item Response Theory (Rating Scale Model)	Winsteps, R

* The level-2 reliability coefficient recommendation is the Spearman-Brown Coefficient.

Limitations

As with any Monte Carlo simulation study, several important limitations exist which may have affected the results. First, the inability to define or apply a theoretical framework from which to interpret the results beyond hypothetical situations is inherent to any simulation study. Second, Monte Carlo simulation procedures are data intensive designs requiring researchers to make numerous consequential decisions regarding data conditions and sampling designs. Choosing the methods of generating mixed and non-normal distributions and their specific characteristics likely constrained generalizability of the results. For example, I originally generated non-normal distributions with a skew = 1.75 and kurtosis = 3.0 as found in the literature; however, this yielded no discernable differences in reliability estimates. Therefore, I increased skew and kurtosis until I was able to detect differences in reliability estimates across data conditions and sampling designs. This threshold was met when skew = 3.0 and kurtosis = 7.0, which represents extremely non-normal data that may not often be found in real-world environments. Next, to generate mixed data distributions, I generated 5 of the 10 items in each data set using a multivariate normal distribution and the other five items using the non-normal data distribution described above. By mixing the distributions at the item level rather than the person level, I put more emphasis on individual items rather than total average scores. Third, to better manage the simulation, I chose to assess only a unidimensional model and to hold the number of items and number of response choices constant. In the single-level and level-1 of two-level models, I fixed reliability estimates to .70, which represents adequate but not excellent reliability. I made these decisions to better reflect real world data scenarios. Each of the decisions I made had consequences on the level of bias in the

reliability coefficients. Finally, Monte Carlo simulations will never capture all of the possible data conditions and sampling designs implemented by applied researchers, limiting the application and generalizability of the results.

Recommendations for Future Research

To better understand relative bias in reliability estimates in multilevel models, future research should include Monte Carlo simulation studies examining polytomous responses within the Rasch RSM framework, not only for person reliability, but to include person separation, and item reliability and separation indices. In addition, responses choices and the number of items should be varied across sample sizes and data distributions. Since 2012, the examination of CTT reliability coefficients in multilevel models has gained momentum and this dissertation adds to that body of literature; however, more research into estimating reliability in single and multilevel models under the umbrella of IRT models, to include RSM and partial credit models is lacking. The results of this dissertation only scratch the surface of how much bias is found in reliability estimates in Rasch RSM models using polytomous response choices and do not address partial credit models or data representing more than one dimension.

Geldhof et al. (2014), Novick and Lewis (1967), Pastore and Lombardi (2014), Rodriguez and Maeda, (2006), Sijtsma (2009), Tavakol and Dennick (2011), Teo and Fan (2013), and Zumbo et al. (2007) argued that Cronbach's α is not the best choice to assess internal consistency in single and multilevel models because (a) it is often underestimated as it represents the lower bound of reliability, (b) the assumption of tau-equivalence is unrealistic, and (c) the assumption of a unidimensional measure is rarely realized. These researchers explained that people continue to use Cronbach's α to report reliability

estimates because they either do not know any better or because it is easy to use as it is readily available in statistical software packages. Prior to conducting the current study for my dissertation, I too felt that Cronbach's α should be replaced by "better" estimates of reliability and that behavioral, educational, and social science researchers would need to accept the inevitability of the call by methodologists to learn another form of reliability estimation. The results of my dissertation show that the average relative bias found in Cronbach's α in single and multilevel models, across data distributions and sample sizes was only slightly higher than polychoric ordinal α and in many cases was less than for polychoric ordinal α. Cronbach's α remains a valid form of assessing reliability, as long as it is understood that reliability estimates may be underestimated and just as with polychoric ordinal α and person reliability, has more bias when distributions are mixed or non-normal. This is good news for applied social researchers because Cronbach's α is readily available in software packages such as SPSS, SAS, and R, better understood than other estimates of reliability across disciplines, and easily interpretable. While perhaps Cronbach's α is not the best choice under every data condition, underlying distribution, and sampling design, it is still a viable method for estimating reliability and should not be deserted for more complicated estimates.

These results are promising since behavioral, educational, and social scientists are often defensive about their use of Cronbach's α yet resistant to using other forms of reliability estimates. Through the process of examining reliability coefficients under various data conditions and sampling designs, I have come to respect the commitment previous researchers have demonstrated in providing applied researchers and clinicians guidelines on the accurate use of reliability estimates in the fields of behavioral,

educational, and social science. I aspire, through the results of this dissertation, to exhibit that same level of commitment to providing meaningful guidance to behavioral, educational, and social science researchers.

REFERENCES

Aiken, L. R. (1983). Number of response categories and statistics on a teacher rating scale. *Educational and Psychological Measurement, 43*(2), 397-401. doi:10.1177/001316448304300209

Allen, I.E. & Seaman, C.A. (2007). Likert scales and data analyses. *Quality Progress*. 40. 64-65.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

Allport, G. W. (1929). The composition of political attitudes. *American Journal of Sociology, 35*, 220-238.

American Educational Research Association. (2014). *Standards for educational and psychological testing: National council on measurement in education, joint committee standards for educational testing*. Ann Arbor, MI: University of Michigan Press.

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42,* 69-81.

Anderson, J., & Gerbing, D. (1984). The effect of sampling error on convergence, improper solutions, and goodness of fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika, 49*(2), 155-173.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561-573.

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park, CA: Sage.

Andrich, D. (2004). Controversy and the Rasch model: a characteristic of incompatible paradigms? *Medical Care*, 42, 1-16.

Annenberg, G. (2012). Annenberg learner. *Measuring accurately.* Retrieved from https://www.learner.org/courses/learningmath/measurement/session2/ part

Asparouhov, T., & Muthén, B. (2006). Multilevel modeling of complex survey data. *Proceedings of the Joint Statistical Meeting in Seattle, WA. American Statistical Association section on Survey Research Methods*, 2718-2726.

Bandelos, L., & Enders, C. (1996). The effects of nonnormality and number of response categories on reliability. *Applied Measurement in Education, 9*(2), 151-160.

Baugh, F. (2002). Correcting effect sizes for score reliability: A reminder that measurement and substantive issues are linked inextricably. *Education and Psychological Measurement, 62*(2), 254-263.

Bearden, W.O., Sharma, S., and Teel, J.E. (1982). Sample size effects on Chi square and other statistics used in evaluating causal models. *Journal of Marketing Research 19*, 425-430.

Beeckman, D., Vanderwee, K., Demarre', L., Paquay, L., Van Hecke, A., & Defloor, T. (2010). Pressure ulcer prevention: Development and psychometric validation of a knowledge assessment instrument. *International Journal of Nursing Studies, 47*(4), 339-410.

Bejar, I. I. (1977). An application of the continuous response level model to personality measurement. *Applied Psychological Measurement, 1*, 509-521.

Bell, B. A., Ferron, J. M., & Kromrey, J. D. (2008). Cluster size in multilevel models: The impact of sparse data structures on point and interval estimates in two-level models. *JSM Proceedings, Section on Survey Research Methods. Alexandria, VA*: American Statistical Association, 1122-1129.

Bentler, P. M. (2006). *EQS 6 Structural equations programs manual*. Los Angeles, CA: University of California at Los Angeles.

Bentler, P. (2009). Alpha, dimension-free and model-based internal consistency reliability. *Psychometrika, 74*, 137-143. doi:10.1007/s11336-008-9001-1

Bernaards C. A., & Jennrich R. I. (2005). Orthomax rotation and perfect simple structure. *Psychometrika, 8,* 585-588.

Biemer, P., Christ, S., & Wiesen, C. (2009). A general approach for estimating scale score reliability for panel survey data. *Psychological Methods, 14*(4), 400-412.

Birnbaum, A. (1957). *Efficient design and use of tests of ability for various decision-making problems*. (58-16 Project no. 7755-23). Randolph Air Force Base, TX: USAF.

Black, R., Yang, Y., Bietra, D., & McCaffrey, S. (2015). Comparing fit and reliability estimates of a psychological instrument using a second order CFA, bi-factor, and essentially tau-equivalent (coefficient alpha) models via Amos 22. *Journal of Psychoeducational Assessment, 33*(5), 451-472.

Bonanomi A., Nai Ruscone M., & Osmetti S. A. (2012). Reliability measurement for polytomous ordinal items: the empirical polychoric ordinal Alpha. *Quad. Statist,* 14, 53-56.

Bonanomi, A., Ruscone, M. N., & Osmetti, S. A. (2013, May). *The polychoric ordinal α measuring the reliability of a set of polytomous ordinal items*. Paper presented at the SIS-International Statistics of Italy, Italy. Retrieved from SIS website #2651

Bond, T. G., & Fox, C. (2014). *Applying the Rasch model: Fundamental measurement in the human science* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Bonito, J. A., Ruppel, E. K., & Keyton, J. (2012). Reliability estimates for multilevel designs in group research. *Small Group research, 43,* 443-469. doi:10.1177/1046496412437614

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*: Springer.

Boomsa, A. (1985). Noncovergence, improper solutions, and starting values in LISREL maximum likelihood estimation. *Psychometrika 50,* 229-242.

Brennan, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice, 38,* 295-317.

Brennan, R. L., & Lee, W. C. (1999). Conditional scale-score standard errors of measurement under binomial and compound binomial assumptions. *Educational and Psychological Measurement, 59*(1), 5-24.

Brown, T. A. (2015). *Confirmatory factor analysis for applied Research* (2nd ed.). New York, NY: Guilford Press.

Brown, W. (1910). Some experimental results in correlation of mental abilities. *British Journal of Psychology, 3,* 296-322.

Browne W. J., & Draper D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics*, *15*, 391-420.

Bryk, A., & Raudenbush, S. W. (1992). *Hierarchical linear models for social and behavioral research: Applications and data analysis methods.* Newbury Park, CA: Sage

Busing F. (1993). *Distribution characteristics of variance estimates in two-level models* . Unpublished manuscript, Leiden University, the Netherlands

Charter R. A. (1997) Confidence interval procedures for retest, alternate-form, validity, and alpha coefficients. *Perceptual and Motor Skills*, 84, 1488-1490

Charter R. A. (1999) Sample size requirements for precise estimates of reliability, generalizability, and validity coefficients. *Journal of Clinical and Experimental Neuropsychology*, 21, 559-566.

Charter, R. A. (2003). A breakdown of the reliability coefficients by test type and reliability method, and the clinical implications of low reliability. *Journal of General Psychology*, *130*(3), 290-304.

Choi, J., Dunlop, M., Chen, J., & Kim, S. (2011). A comparison of different approaches for coefficient α for ordinal data. *Journal of Educational Evaluation*, *24*(2), 485-506.

Churchill G. A., Jr., & Peter J. P. (1984) Research design effects on the reliability of rating scales: A meta-analysis. *Journal of Marketing Research*, 21, 360-375

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284-290.

Clark, P. (2008). When can group level clustering be ignored? Multilevel models versus single level models with sparse data. *Journal of Epidemiology and Community Health, 62*, 752-458.

Coe, R. (2000). *What is an 'effect size?' A guide for users*. Retrieved from http://www.cemcentre.org/ebeuk/research/effectsize/ESguide.htm.

Coe, R. (2002, September 12-14). *It's the effect size, stupid: What effect size is and why it is important.* Paper presented at the British Educational Research Association Annual Conference, Exeter, England. 12-14 September, 2002.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York, NY: Academic Press.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York, NY: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Converse, J. M. (2009). *Survey research in the United States: Roots and emergence 1890-1960*. New Brunswick, NJ: Transaction Publishers.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98-104.

Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: Holt, Rinehart, and Winston.

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

Culligan, B. (2013). Item response theory, reliability, and standard error. *Research in Item Response Theory*. 6, 211-217.

Culpepper, S. A. (2013). The reliability and precision of total scores and IRT estimates as a function of polytomous IRT parameters and latent trait distribution. . *Applied Psychological Measurement, 37*(3), 201-225. doi:10.1177/0146621612470210

Daher, A. M., Ahmad, S. A., Winn, T., & Selamat, M. I.(2015). Impact of rating scale categories on reliability and fit statistics of the Malay Spiritual well-being Scale using Rasch analysis. *Malays Journal of Medical Science, 22*(3), 48-55.

Davidson, I., Cooper, R., & Bullock, A. (2010). The objective structured public health examination: A study of reliability using multilevel analysis. *Medical Teacher, 32*(7), 582-585.

Dedrick, R. F., & Greenbaum, P. E. (2010), Multilevel confirmatory factor analysis of a scale measuring interagency collaboration of children's mental health agencies. *Journal of Behavioral Disorders, 19*, 27-40.

DeMars, C. (2010). *Item response theory: Understanding statistics*. New York, NY: Oxford Press.

Dempster, A. P., Rubin, D. B., & Tsutakawa, R. K. (1981). Estimation in covariance components models. *Journal of the American Statistical Association, 76*(374), 341-353.

Dick, W., & Hagerty, N. (1971). *Topics in measurement: Reliability and validity*. New York, NY: McGraw-Hill.

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling, 9*(3), 327-346.

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology*, *47*, 309-326.

Draxler C. (2010). Sample size determination for Rasch model tests. *Psychometrika, 75*, 708-724. doi:10.1007/s11336-010-9182-4

Duhachek, A., Coughlan, A. T., & Iacobucci, D. (2005). Results on the standard error of coefficient alpha index of reliability. *Marketing Science, 24*(2), 294-301. doi:10.1287/mksc.1040.0097

Durlak, J. A. (2009). How to select, calculate, and interpret effect sizes. *Journal of Pediatric Psychology, 34*(9), 917-928.

Ekström, J. (2009). Contributions to the theory of measures of Association for ordinal variables, digital comprehensive summaries of Uppsala Dissertation from the Faculty of Social Sciences, n. 50. ACTA Universitatis Upsaliensis, Uppsala.

Ekström, J. (2010). On the relation between the polychoric correlation coefficient and Spearman's rank correlation coefficient. *Department of Statistics Papers. Retrieved from the UCLA* website: https://escholarship.org/uc/item/7j01t5sf

Emberson, S. P., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Enders, C. K. (2008). A note on the use of missing auxiliary variables in full information maximum likelihood-based structural equation models. *Structural Equation Modeling, 15*(1), 434-448.

Englehard, G. (2014). *Item response theory models for rating scale data.* Hoboken, NJ: John Wiley & Sons.

ETS. (1995). *TOEFL test and score data summary 1995*. Retrieved from http://ets.org/TOEFL

Fan, X. (1998). Item response theory and Classical Test Theory: An empirical comparison of their item/person characteristics. *Education and Psychological Measurement, 58,* 357-381.

Feldt, L. S. (1990). The sampling theory for intraclass reliability coefficient. *Applied Measurement in Education, 3,* 361-367.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. *Educational Measurement,* 3rd ed., 105-146.

Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement, 11*, 93-103.

Finney, S. J., & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 269-314). Greenwich, CT: Information Age Publishing.

Fisher, C. R. (2014). A pedogogic demonstration of attenuation of correlation due to measurement error. *Spreadsheets in Education, 7*(1-Article 4).

Fitzmaurice, G. (2002). Measurement error and reliability. *Nutrition, 18*(1), 112-114. doi:10.1016/S0899-9007(01)00624-4

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*(4), 466-491.

Frost, J. (2015). *Choosing between a nonparametric test and a parametric test*. Retrieved from http://blog.minitab.com/blog/adventures-in-statistics-2/choosing-between-a-nonparametric-test-and-a-parametric-test

Furr, M., & Bacharah, V. R. (2014). *Psychometrics: An introduction* (2nd ed.). Thousand Oaks, CA: Sage Publishing.

Gaberson, K. B. (1997). Measurement reliability and validity. *AORN Journal, 66*(6), 1092-1094.

Gadermann, A. M., Gruhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment Research & Evaluation, 17*(3), 1-13.

Geldhof, G. J., Preacher, K. J., & Zyphur, m. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Journal of Psychological Methods, 19*(1), 72-91. doi:10.1037/a0032138

Gerhart, B., Wright, P., Mahan, M. C., & Snell, S. A. (2000). Measurement error in research on human resources and firm performance: how much error is there and how does it influence effect size estimates. *Personnel Psychology, 53*(4), 803-834.

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference 11.0 update* (4th ed.). Boston, MA: Allyn & Bacon

Gliner, J. A., Morgan, G. A., & Harmon, R. J. (2001). Measurement reliability. *Journal of American Academy of child & adolescent psychiatry, 40*(4), 98-102.

Goldman, A., & Mitchell, D. F. (2008). Directory of unpublished experimental mental measures. *American Psychologial Association, 11*(9), 120-129.

Goodman, L. A., & Kruskal, W. H. (1979). *Measures of association for cross classifications*. New York, NY: Springer-Verlag.

Green, K. E., & Frantom, C. G. (2002). *Survey development and validation with the Rasch model*. Paper presented at the International Conference on Questionnaire Development, Evaluation, and Testing, Charleston, SC.

Green, S. B., Akey, T. M., Fleming, K. K., Hershberger, S. L., & Marquis, J. G. (1997). Effect of the number of scale points on chi-square fit indices in confirmatory factor analysis. *Structural Equation Modeling, 4*, 108-120

Guttman, L. (1950). The basis for scalogram analysis. *The American Soldier, IV* (Measurement and Prediction). New York, NY: Wiley

Guttman, I. (1967). The use of the concept of future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society, 29*(1). 83-100.

Haldago-Tello, F. P., Chacón-Moscoso, S., Barbero-García, I., & Vila-Abad, E. (2008). Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality and Quantity, 44*(153), 26-34.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of Classical Test Theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 47-53.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and itenis. *Applied Psychological Measurement, 9*(2), 139-164.

Hellman, C. M., Fuqua, D. R.. & Worely, J. (2006). A reliability generalization study on the survey of perceived organizational support. *Educational and Psychological Measurement, 66*(4), 631-642.

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

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). reliability methods. *Educational and Psychological Measurement, 60*, 523-531.

Hox, J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Erlbaum

Huang, F. L., & Cornell, D. G. (2016). Multilevel factor structure, concurrent validity, and test-retest reliability of the high school teacher version of the authoritative school climate survey. *Journal of Psychoeducational Assessment, 34*(6), 536-549.

Huberty, C. J. (2002). A history of effect size indices. *Educational and Psychological Measurement, 62*(2), 197-226.

Hutchinson, S. R., & Olmos, A. (1998). Behavior of descriptive fit indexes in confirmatory factor analysis using ordered categorical data. *Structural Equation Modeling, 5*(4). 344-364.

Jaccard, J., & Turrisi, R. (2003). *Interaction effects in multiple regression (Volume 7).* New York, NY: Sage Publishing.

Jackson, D. L. (2003). Revisiting sample size and number of parameter estimates: Some support for the N;q hypothesis. *Structural Equation Modeling, 10*(1), 128-141.

Jenkins, G. D., & Taber, T. D. (1977). A Monte Carlo study of factors affecting three indices of composite scale reliability. *Journal of Applied Psychology, 62* (4), 392-398

Johns, R. (2010). *Likert items and scales*. Retrieved from Survey Question Bank website: http://www.surveynet.ac.uk/ sqb/datacollection/likertfactsheet.pdf

Jöreskog K. G., & Sörbom, D. (1986). *LISREL VI: Analysis of linear structural relationships by maximum likelihood, instrumental variables, and least squares methods*. Mooresville, IN*:* Scientific Software, Inc.

Kahn, M. I. (2014). Recovery and stability of item parameter and model fit across varying sample sizes and test lengths o Rasch analysis with small sample. *Social Science International 30*(5), 43-60.

Kamakura, W. A., & Balasubramanian, S. (1989). Tailored interviewing: An application of item response theory for personality measurement. *Journal of Personality Assessment, 53,* 502-519.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38*(1), 79-93.

Kanyongo G. Y., Brook, B. P., Kyei-Blankson, L. K., & Gocmen, G. (2007). Reliability and statistical power: How measurement fallibility affects ower and required sample sizes for several parametric and nonparametric statistics. *Journal of Modern Applied Statistical Methods, 6*(1), 81-90.

Karabatsos, G. (2000). *Principles of scale development*. Paper presented in a meeting of the Medical Decision-Making Committee, Louisiana State University Health Sciences Center, Baton Rouge, LA.

Kim, S., & Feldt, L. S. (2010). The estimation of the IRT reliability coefficient and its lower and upper bounds with comparisons to CTT reliability statistics. *Asia Pacific Education Review, 11*(2), 179-188.

Kline, J. B. (1999). Review of psychometric theory, Nunnally and Bernstein: 1994. *Journal of Psychoeducational Assessment*, *17*, 275-280.

Kline, J. B. (2005). *Psychological testing: A practical approach to design and evaluation*. Thousand Oaks, CA: Sage.

Kline, R. (2009). *Becoming a behavioral science researcher: A guide to producing research that matters*. New York, NY: Guilford Press.

Kline, R. (2014). *Principles and practice of structural equation models* (4th ed). New York, NY. Guilford Press.

Ko, Y.-H. (1994). A search for a better Likert point-scale for Mental Health Questionnaires. *Psychological Testing*, 41, 55-72.

Kolen, M.J. & Tong, Y. (2009) National council on measurement in education. Iowa City, IA. University of Iowa Press.

Kopalle, P. K., & Lehmann, D. R. (1997). Alpha inflation? The impact of eliminating scale items on Cronbach's alpha. *Organizational, Behavioral and Human Decision Processes, 70*(3), 189-197.

Lance, C. E., & Vandenberg, r. J. (2010). *Statistical and methodological myths and urban legends: Doctrine, verity, and fable in the organizational and social sciences*. New York, NY: Sage.

Lawley, D. N. (1943). On problems connected with item selection and test construction. *Proceedings for the Royal Society of Edinburgh, 61*, 273-287.

Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 4*(2), 73-79.

Lewis, C. (2007). Classical Test Theory: In C.R. Rao and S Sinharay; Handbook of statistics. *Psychometrics, 26*, 29-43.

Libarkin, J. C., & Anderson, S. W. (2005). Assessment of learning in entry-level geoscience courses: results from the Geocience Concept Inventory. *Journal of Geoscience in Education, 53*(4), 394-401.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22*, 5-55.

Linacre. J. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7*(4), 328-330.

Linacre, J. M. (2004). *Estimation methods for Rasch measures*. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 226-257). Maple Grove, MN: JAM Press.

Linacre, J. M. (2012). Empirical multidimensionality or imposed unidimensionality? *Rasch Measurement Transactions, 26*(2), 1372.

Linacre, J. M. (2014). False precision: Too many rating scale categories. *Rasch Measurement Transactions, 28*(2), 1463-1464.

Linacre, J. M. (2017). Measurement decision theory from a Rasch perspective. *Rasch Measurement Transactions, 31*(1), 1618-1621.

Lindquist, M. M. (1989). *Results from the fourth mathematics assessment of the National Assessment of Education Progress.* Paper presented at the National Council of Teachers in Mathematics, Reston, VA.

Lissitz, R. W., & Green, S. B. (1975). Effect of the number of scale points on reliability: A Monte Carlo approach. *Journal of Applied Psychology, 60(*1), 10-13.

Little, J. (2013). Multilevel confirmatory ordinal factor analysis of the Life Skills Profile-16. *Psychological Assessment, 25*(3), 810-825.

Liu, Y., Wu, D., & Zumbo, B. (2009). The impact of outliers on Cronbach's coefficient alpha estimates of reliability: Ordinal/rating scale item responses. *Educational and Psychological Measurement, 72*(1), 5-21.

Lord, F. M. (1952). The relation of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika, 17*(2), 181-194.

Lord, F. M. (1953). On the statistical treatment of football number. *American Psychologist, 8.* 750-751.

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

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology, 1*, 1-27.

Ludtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel covariate model. *Psychological Methods*, *13* (3) 203-229.

Luke, D. A. (2004). *Multilevel modeling*. Thousand Oaks, CA: Sage.

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*, 86-92.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996) Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1,* 130-149.

Maddox, T. (Ed.). (1997). *Tests: A comprehensive reference for assessments in psychology, education, and business* (4th ed.). Austin, TX: Pro-Ed publications.

Marais, I., & Andrich, D. (2008) Effects of varying magnitude and patterns of local dependence in the unidimensional Rasch model. *Journal of Applied Measurement. 9*(2) ,105-124.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Masters, G. (1988). The analysis of partial credit scoring. *Applied Measurement in Education*, 1, 279-297.

Matell, M. S., & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items? Study I: Reliability and validity. *Educational and Psychological Measurement, 31*(3), 657-674.

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods, 11*(4), 344-362.

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

McIver, J. P., & Carmines, E. G. (1981). *Unidimensional Scaling*. Thousand Oaks, CA: Sage.

Messick, S. (1989). *Validity*. New York, NY: Macmillan.

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*(3), 241-256

Miller, M. J. (2004). Reliability and validity. *RES 600, Western International University, 1*(3), 122-124.

Moore, H. T. (1925). Innate factors in radicalism and conservatism. *Journal of Abnormal and Social Psychology, 20,* 234-244.

Moss, P.A. (1994). The meaning and consequences of "reliability." *Journal of Educational and Behavioral Statistics, 29*(2), 245-249.

Murphy, G. (1929). *An historical introduction to modern psychology*. New York, NY: Harcourt, Brace, and Company.

Murphy, G., & Murphy, L. B. (1931). *Experimental social psychology*. New York. NY: Harper.

Murphy, L., Impara, J., & Plake, B. (1999). *Tests in Print*. Highland park, NJ: Gryphon Press.

Muthén, B. (1981). Factor analysis of dichotomous variables: American attitudes toward abortion. In D. J. Jackson & E. F. Borgatta (Eds.), *Factor analysis and measurement in sociological research: A multi-dimensional perspective*. London, England: Sage.

Muthén, B. (1983). Latent variable structural equation modeling with categorical data. *Journal of Econometrics*, *22,* 48-65.

Muthén, B. (1987). Response to Freedman's critique of path analysis: Improve credibility by better methodological training. Journal of Educational Statistics, (12), 178-184.

Muthén, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, 54, 557-585.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research, 22*, 376-400. doi:10.1177/0049124194022003006

Muthén, B., & Asparouhov, T. (2009). Multilevel regression mixture analysis. *Journal of the Royal Statistical Society,* 172, 639-657.

Muthén, B., & Asparouhov, T. (2011). Beyond multilevel regression modeling: Multilevel analysis in a general latent variable framework. In J. Hox & J.K. Roberts (Ed.s), *Handbook of advanced multilevel analysis* (pp. 15-40). New York, NY: Taylor and Francis.

Muthén, B., & Kaplan D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. British Journal of Mathematical and Statistical Psychology, 38, 171-189. Retrieved from http://dx.doi.org/10.1111/j.2044-8317.1985.tb00832.x

Muthén, B., & Muthén, L. K. (2000). Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and Experimental Research, 24*(6), 882-891. doi:10.1097/00000374-200006000-00020

Muthén, B., & Satorra A. (1995). Complex sample data in structural equation modeling. In P. V. Marsden (Ed.), *Sociological methodology* (pp. 267-316). Oxford, England: Blackwell

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(4), 599-620. doi:10.127/S15328007SEM0904_8

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Science Education*. 15, 625-632

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometricka, 32(*1), 1-13.

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

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: Erlbaum.

Nunnally, J. C. (1982). *Reliability of measurement*. *Encyclopedia of educational research*. New York, NY: Macmillan.

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

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*(4), 443-460.

Oppenheim, A. N. (1986). *Questionnaire design and attitude measurement.* Glencoe, IL: The Free Press of Glecoe.

Oshima, T. C., & McCarty, F. (2015). *Factor analysis of variance. Statistically significant interactions: What's the next step?* [Scholarly project]. In *Georgia State University Scholarship*, *10*(3), 244-256.

Ostini, R., & Nering, M. (2006). *Polytomous item response theory models*. Thousand Oaks, CA: Sage.

Pastor, D. A. (2003). The use of multilevel item response theory modeling in applied research: An illustration. *Applied Measurement in Education, 16*(3), 223-243. doi:10.1207/s15324818ame1603

Pastore, M., & Lombardi L. (2014). The impact of faking on Cronbach's alpha for dichotomous and ordered rating scores. *Qual. Quant, 48*(2), 1191-1211. doi:10.1007/s11135-013-9829-1

Pearson, K. (1900). Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London, 195*(Series A), 1-47.

Pearson, K. (1907). *On further methods of determining correlation. Draper's Company Research Memoirs Biometric Series, IV*. London, England: Cambridge University Press.

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

Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch model as additive conjoint measure. *Applied Psychological Measurement, 3*(2), 237-255.

Peterson, R. A. (1994) A meta-analysis of Cronbach's coefficient alpha. *Journal of Consumer Research*, 21, 381-391

Pickering, S. (2001). Coefficient alpha and reliability of scale scores. *Applied Psychological Measurement, 32(*5). 122-136.

Preacher, K. J., Wichman, A. L., MacCallum,R. C., & Briggs, N. E. (2012). Latent growth curve modeling. *Structural Equation Modeling. 19(*1), 152-155.

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychological, 104*(1), 1-15. doi:10.1016/s0001-6918(99)00050-5

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danish Institute for Educational Research.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *The Danish Yearbook of Philosophy*, 14, 58-93.

Raudenbush, S. W. (1993). Hierarchical linear models as generalizations of certain common experimental design models. In L. Edwards (Ed.). Applied analysis of variance in behavioral science(pp. 459-496, Chapter 13). New York, NY: Marcell Decker.

Raudenbush, S. W. (1994). Analyzing effect sizes: Random effects models. In H. Cooper, L. V. Hedges (Eds.). The handbook of research synthesis (pp. 302-32, Chapter 20). New York, NY: Russell Sage Foundation.

Raudenbush, S. W., & Bryk, A. S. (1994). *Hierarchical linear models. International Encyclopedia of Education: Research and studies* (2nd ed.). Oxford, England: Pergamon Press.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage

Raudenbush, S. W., Johnson, C., & Sampson, R. J. (2003). A multivariate, multilevel Rasch model for self-reported criminal behavior. *Sociological Methodology, 33*(1), 169-211.

Raykov, T. (1997a). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21*, 173-184

Raykov, T. (1997b). Scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components. *Multivariate Behavioral Research, 32*, 329-353.

Raykov, T. (2010). Proportion of third-level variance in multilevel studies: A note on an int3erval estimation procedure. *British Journal of Mathematical and Statistical Psychology, 64*, 38-52.

Raykov, T., & Penev, S. (2010). Estimation of reliability coefficients in two-level designs via latent variable modeling. *Structural Equation Modeling, 17*, 629-641.

Reeve, B. B., & Fayers, P. (2005). *Applying item response theory modeling for evaluating questionnaire items and scale properties*. Oxford, England: Oxford University Press.

Revelle, W. (2011). *R psych: Procedures for personality and psychological research*, Northwestern University, Evanston, Illinois, USA

Richardson, M. W. (1936). The relationship between difficulty and the differential validity of a test. *Psychometrika, 1*, 33-49.

Rodriguez, M. C., & Maeda, Y. (2006). Meta-analysis of coefficient alpha. *Journal of Psychological Methods, 11*(3), 306-322.

Roiser, M. (1996). Consensus, attitudes, and Guttman scales. *Empirical Investigations in Social Representations, 5*, 122-128.

Rosenberg, B. (1973). Random coefficients models: The analysis of a cross section of time series by stochastically convergent parameter regression. *Annals of Economic and Social Measurement, 2*(4), 1033-1042.

Rosenthal R., & Rosnow R. L. (1991) *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York, NY: McGraw-Hill.

Rothke, S. (2010). *Basic measurement and statistics*. Chicago, IL. Kendall Hunt.

Roxy, P., Olson, C., & Davore, J. L. (2011). *Introduction to statistics and data analysis* (5th ed). New York, NY. Brooks Cole.

Rust, J., & Golombok, S. (2008). *Modern psychometrics: The science of psychological assessment* (3rd ed.). London, England: Routledge.

Rutkowski, L., & Svetina, D. (2013). Assessing the hypothesis of measurement Invariance in the context of large-scale international surveys. *Educational and Psychological Measurement, 74*(1), 31-57. doi:10.1177/0013164413498257

Salkind, N. J. (2010). *Encyclopedia of research design*. Los Angeles: SAGE.

Samejima, F. (1977). The use of the information function in tailored testing. *Applied Psychological Measurement,* 1, 233-247.

Samejima, F. (1979). Convergence of the conditional distribution of the maximum likelihood estimate, given latent trait, to the asymptotic normality: Observations made through the constant information model (*Office of Naval Research Report, 79-3*). Arlington, VA: Office of Naval Research.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8,* 350-353. doi:10.1037/1040-3590.8.4.350

Segall, D. O. (1994). The reliability of linearly equated tests. *Psychometrika, 59,* 361-375

Serfling, R. (2010). Equivariance and invariance properties of multivariate quantile and related functions, and the role of standardization. *Journal of Nonparametric Statistics, 22*(7), 915-936. doi:10.1080/10485250903431710

Shavelson, R. J., & Webb, N. M. (1991). *Measurement methods for the social sciences: Generalizability theory*. Newbury Park, CA. Sage.

Sheng, Y., & Sheng, Z. (2012). Is coefficient alpha robust to non-normal data? *Frontiers in Psychology, 3*(34), 1 -15. doi:10.3389/fpsyq.2012.00034

Shevlin, M., Miles, J., Davies, M., & Walker, S. (2000). Coefficient alpha: A useful indicator of reliability? *Personality and Individual Differences, 28*(2), 229-237.

Sideridis, G. D. (1999). Examination Of The Biasing Properties Of Cronbach Coefficient Alpha Under Conditions Of Varying Shapes Of Data Distribution: A Monte Carlo Simulation. *Perceptual and Motor Skills, 89*(7), 899. doi:10.2466/pms.89.7.899-902

Sijtsma, K. (2009). On the use, misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*(1), 107-120.

Smith, R. M. (1993). Guessing and the Rasch Model. *Rash Transactions of Rasch Measurement SIG American Educational Research Association Measurement: , 6*(4), 262-263.

Snijders, T. A. (2005). *Power and sample size in multilevel linear models: Encyclopedia of statistics in behavioral science*. New York, NY: John Wiley and Sons. doi:10.1002/0470013192.bsa492

Snijders T. A. B., & Bosker R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

Snyder, C. R. (1994). *The psychology of hope: You can get there from here*. New York, NY: The Free Press.

Spearman, C. (1904). General intelligence: Objectively determined and measured. *American Journal of Psychology, 15*, 201-293.

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

Stevens, S. S. (1946). On the theory of scale measurements. *Science, 103*(2684), 677-680. doi:10.1126/science.103.2684.677.

Stevens, S. S. (1951). Mathematics, measurement and psychophysics. *Handbook of Experimental Psychology*, 1-49.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2*, 53-55.

Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification, 30*(2), 251-275. doi:10.1007/s00357-013-9129-4

Teo, T., & Fan, X. (2013). Coefficient alpha and beyond: Issues and alternatives for educational research. *The Asia-Pacific Education Researcher, 22*(2), 209-213.

Thissen, D., & Wainer, H. (Ed.s). (2001). *Test scoring*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement, 54*, 837-847.

Thompson, B. (Ed.). (2003). *Score reliability: Contemporary thinking on reliability issues*. Thousand Oaks, CA: Sage

Thompson, B. (2007). *Foundations of behavioral statistics. An insight-based approach*. New York, NY: Guilford Press.

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

Thorndike, E. L. (1904). *An introduction to the theory of mental and social measurements*. New York, NY: Teachers College University.

Thorndike, E. L. (1913). Educational Diagnosis. *Science, 37*(943), 133-142. doi:10.1126/science.37.943.133

Thorndike, E. L. (1918). Individual differences. *Psychological Bulletin, 15*(5), 148-159. doi:10.1037/h0070314

Thorndike, E. L. (1919). *Educational psychology: Briefer course*. New York, NY, Columbia Press. doi:10.1037/10608-000

Thorndike, R. M., & Thorndike-Christ, T. (2010). *Measurement and evaluation in psychology and education* (8th ed.). Boston, MD: Pearson.

Thurstone, L. L. (1924). *The nature of intelligence*. London, England: Routledge.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*, 529-554.

Townsend, P. D., Christensen, M. G., Kreiter, C. D., & ZumBrunnen, J. R. (2010). Investigating the use of written and performance-based testing to summarize competence on the case management component of the NBCE Part IV-national practical examination. *Teaching and Learning in Medicine, 23*(1), 16-21.

Traub, R. E. (1997). Classical Test Theory in historical perspective. *Educational Measurement: Issues and Practice, 16*(4), 14-20. doi:10.1111/j.1745-3992.1997.tb00604.x

Ulf, B., & Lehmann, D. (2015). On the limits of research rigidity: The number of items in a scale. *Marketing Letters, 26*(3), 257-260.

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

van der Eijk, C., & Rose, J. (2015). Risky business: Factor analytics of survey data-assessing the probability of incorrect dimensionalisation. *Plos One 10*(3), 1-20.

Van der Leeden R., & Busing F. (1994). *First iteration versus IGLS RIGLS estimates in two-level models: A Monte Carlo study with ML3*. Unpublished manuscript, Leiden University, the Netherlands

Van der Leeden, R., Busing, F., & Meijer, E., (1997). *Applications of bootstrap methods for two-level models*. Unpublished paper, Leiden University, the Netherlands

van Zyl, J. M., Neudecker, H., & Nel, D. G. (2000). On the distribution of the maximum likelihood estimator of Cronbach's alpha. *Psychometrika, 65*(3), 271-280.

Weiner, J. (2007). *Measurement: Reliability and validity measures*. Retrieved from http://ocw.jhsph.edu/courses/ HSRE/PDFs/HSRE_lect7_weiner.pdf

Weng, L. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement, 64*(6), 956-972. doi:10.1177/0013164404268674

Wilkinson, L. (1999). Wilkinson, L. and the task force on statistical inference. *American Psychologist, 54*, 594-604. doi:10.1037/0003-066X.54.8.594

Wright, B. D. (1967). Sample-free test calibration and person measurement. *Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, NJ*.

Wright, B. D. (1984). The essential process in a family of measurement models. *Psychometrika, 49*(4), 529-544.

Wright, D. B., Gaskell, G. D., & O'Muircheartaigh, C. A. (1997). How response alternatives affect different kinds of behavioural frequency questions. *British Journal of Social Psychology, 36*, 443-456.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: Mesa Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA press.

Wright, B. D., & Stone, M. H. (1999). *Measurement essentials* (2nd ed.). Wilmington, DE: Wide Range.

Wu, A. D., & Zumbo, B. D. (2008). Understanding and using mediators and moderators. *Social Science Research, 87*, 367.

Yang, Y., & Green, S. B. (2011). Coefficient alpha: A reliability coefficient for the 21st century? *Journal of Psychoeducational Assessment, 29*(4), 377-392.

Yurdugul, H. (2008). Minimum sample size for Cronbach's coefficient alpha: A Monte-Carlo study. *Hacettepe Üniversitesi Journal of Education, 35*, 397-405.

Zeller, R. A., & Carmines, E. G. (1980). *Measurement in the social sciences: The link between theory and data*. New York, NY: Cambridge Press.

Zhang, O. (2010). *Polytomous IRT or testlet model: An evaluation of scoring models in small testlet size situations* (Unpublished master's thesis). University of Florida, Gainsville, FL.

Ziegler, M., Poropat, A., & Mel, J. (2014). Does length of a questionnaire matter? Expected and unexpected answers from generalizability theory. *Journal of Individual Differences, 35*(4), 250-261.

Zimmerman, D. W., & Williams, R. H. (1986). Note on reliability of experimental measures and the power of significance tests. *Psychological Bulletin, 100*(1), 123-124.

Zimmerman, D. W., Williams, R. H., & Zumbo, B. D. (1993). Reliability of measurement and power of significance tests based on differences. *Applied Psychological Measurement, 17*, 1-9.

Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal versions of coefficient α and theta for Likert rating scales. *Journal of Modern Applied Statistical Methods, 6*(1/4), 21-29.

APPENDIX A

GENERATING MULTIVARIATE DISTRIBUTIONS FOR CROBACH'S ALPHA

GENERATING MULTIVARIATE DISTRIBUTIONS FOR CROBACH'S ALPHA

R-code for a single level model (normal distribution) and Cronbach's alpha. Note that for this dissertation, the code allows for changes to sample size and type of distribution. The code below is for *n* = 300, normal distribution, 10 items, 5 response choices

```
##############################################################################
Pearson Correlation Matrix
##############################################################################
# stuff from last time: input parameters ----loading the packages
library(psych)
library(MASS)
library(psy)
library(MBESS)
library(lavaan)
# pop of 100,000, multivariate normal (7 items) (resampling from this)
# mean of 0 (for all items), correlation between all items is .7
#Pearson Correlation Matrix
# 3 different
# Load a library with a function that can generate multivariate normal data
# This is the function that will generate multivariate random normal data
#?mvrnorm

# Setting population parameters ###
##
# Only change these numbers!
set.seed(1842)
n <- 300 #setting population
sample.size <- 300 #for each iteration
iterations <- 1000 #number of cronbachs alphas we want to find
number.of.items <- 10
sd.of.item <- 1 #within item sd, diagonals of correlation matrix
des.alpha <- .7 #enter the desired alpha level
cor <- des.alpha / (des.alpha + number.of.items - des.alpha*number.of.items) #se
item.min <- 1
item.max <- 5
mean <- mean(c(item.min,item.max))+.5 # add .5 so that you're in the middle of truncated
solutions
###

# Generating multivariate parameters from given input ====
# create the vector of means
vec.means <- rep(mean, number.of.items)

# set inter-item correlation matrix
cor.mat <- matrix(cor, ncol=number.of.items, nrow=number.of.items)
```

```
# set diagonals of correlation matrix to 1
diag(cor.mat) <- sd.of.item

# Generate multivariate normal DATA for Population ====
# setting data as "population"
population <- trunc(mvrnorm(n=n, mu=vec.means, Sigma=cor.mat))
# set minimum acceptable value to the "item.min"
population <- pmax(population, 1)
# set max acceptable value to the "item.max"
population <- pmin(population, 5)
#check data
Population

#Calculating Cronbach's alpha -
cronbach(population)

# draw histogram of item responses -
# dev.off()  # clear plots (if there are any)
# par(mfrow=c(2,4))   # Sets a "plot matrix", so that we can see all 7 plots at the same time
#
# for(i in 1:ncol(population))
# {
#   hist(population[,i], breaks=(item.min:(item.max+1)-.5), main=c("Item",i))
# }

# Actually Sampling From Population ####
#### getting the hang of it
#set.seed(1842) #for repeatability
ids <- sample(x=n, size=sample.size, replace=TRUE) #get identifiers of those in the sample
sam <- population[ids,] #find the people in the population with the identifiers selected (with all
columns)
#cronbach(sam)$alpha

#making the loop
output <- numeric(length=iterations) #NOTE: MISSING IS 0

for(i in 1:iterations){
  ids <- sample(x=n, size=sample.size, replace=TRUE)
  sam <- population[ids,]
  output[i] <- cronbach(sam)$alpha
}

#output

summary(output)
hist(output)

#creating CI for the mean
mean(output) + qnorm(c(.025,.975)) * sqrt(var(output)/length(output))
```

```
# Graphing the alphas
dev.off() #To clear out current plots
hist(output)

# Build CI around cronbachs alpha ----
#using MBESS to get confidence interval
#ci.reliability
#ci.reliability(data=population, type="alpha")

# Error Checking ----
#Checking that the inter-item correlations are close to the desired inter-item correlations
#cor.mat
#cor(population)
#check that item range is from 1 to 7 (nothing greater or less than)
#range(population)
#table(population)

#checking normality
#plot
#hist(population, freq=FALSE) #freq=FALSE gives density instead of frequency
#curve(dnorm(x, mean=2.75)*2, add=TRUE) #2.75 isn't the mean, it is the middle of the bin

#write.csv(population, "c:/")
# Getting the observed lphas into excel
#filepath <- paste0("C:/Users/karen/Documents/KT ASUS/Documents/ASRM/Actual
Dissertation/SNn30.csv"),"/RC", item.max - item.min + 1, "N", sample.size, "I", number.of.items,
".csv")
write.csv(population,"C:/Users/karen/Documents/DISDATA/IRT300norm_date.csv",
row.names=FALSE)
#write.csv(output,"C:/n200i15rc7Nprac.csv")

output
sd(output)
```

APPENDIX B

GENERATING MULTIVARIATE DISTRIBUTIONS FOR POLYCHORIC ORDINAL ALPHA

GENERATING MULTIVARIATE DISTRIBUTIONS FOR POLYCHORIC ORDINAL ALPHA

```
##############################################################################
Polychoric matrix USE cor="poly" 10 items normal
##############################################################################
# stuff from last time: input parameters ----
library(psych)
library(MASS)
library(psy)
library(MBESS)
library(lavaan)
# pop of 100,000, multivariate normal (7 items) (resampling from this)
# mean of 0 (for all items), correlation between all items is .7
# 3 different
# Load a library with a function that can generate multivariate normal data
# This is the function that will generate multivariate random normal data
#?mvrnorm

# Setting population parameters ###
##
# Only change these numbers!
sample.size <- 300 #for each iteration
iterations <- 1000 #number of cronbachs alphas we want to find
number.of.items <- 10
sd.of.item <- 1 #within item sd, diagonals of correlation matrix
des.alpha <- .7 #enter the desired alpha level
cor <- des.alpha / (des.alpha + number.of.items - des.alpha*number.of.items) #se
item.min <- 1
item.max <- 5
mean <- mean(c(item.min,item.max))+.5 # add .5 so that you're in the middle of truncated
        solutions
# Generating multivariate parameters from given input ====
# create the vector of means
vec.means <- rep(mean, number.of.items)
```

```
cor="poly"
# set inter-item correlation matrix
#cor.mat <- matrix(cor, ncol=number.of.items, nrow=number.of.items)
polycor.mat <-PMat(cor,ncol=number.of.items, nrow=number.of.items)

# set diagonals of correlation matrix to 1
diag(cor.mat) <- sd.of.item
cor="poly #to create a polychoric matrix

# Generate multivariate normal DATA for Population ====
# setting data as "population"
population <- trunc(mvrnorm(n=n, mu=vec.means, Sigma=cor.mat))
# set minimum acceptable value to the "item.min"
population <- pmax(population, item.min)
# set max acceptable value to the "item.max"
population <- pmin(population, item.max)
#check data
population
#calculate polychoric ordinal alpha
cronbach(population)#polymat
# draw histogram of item responses -
# dev.off()  # clear plots (if there are any)
# par(mfrow=c(2,4))   # Sets a "plot matrix", so that we can see all 7 plots at the same time
#
# for(i in 1:ncol(population))
# {
#   hist(population[,i], breaks=(item.min:(item.max+1)-.5), main=c("Item",i))
# }

# Actually Sampling From Population ####
#### getting the hang of it
#set.seed(1842) #for repeatability
ids <- sample(x=n, size=sample.size, replace=TRUE) #get identifiers of those in the sample
sam <- population[ids,] #find the people in the population with the identifiers selected (with all
        columns)
```

```
cronbach(sam)$alpha #polychoric since using polymath

#making the loop
output <- numeric(length=iterations) #NOTE: MISSING IS 0

for(i in 1:iterations){
  ids <- sample(x=n, size=sample.size, replace=TRUE)
  sam <- population[ids,]
  output[i] <- cronbach(sam)$alpha
}

#output

summary(output)
hist(output)

#creating CI for the mean
mean(output) + qnorm(c(.025,.975)) * sqrt(var(output)/length(output))



# Graphing the alphas
dev.off() #To clear out current plots
hist(output)

# Build CI around cronbachs alpha ----
#using MBESS to get confidence interval
#ci.reliability
#ci.reliability(data=population, type="alpha")

# Error Checking ----
#Checking that the inter-item correlations are close to the desired inter-item correlations
#cor.mat
#cor(population)
#check that item range is from 1 to 5 (nothing greater or less than)
#range(population)
```

```
#table(population)


#checking normality

#plot

#hist(population, freq=FALSE) #freq=FALSE gives density instead of frequency

#curve(dnorm(x, mean=2.75)*2, add=TRUE) #2.75 isn't the mean, it is the middle of the bin


#write.csv(population, "c:/")

# Getting the observed lphas into excel

#filepath <- paste0("C:/Users/karen/Documents/KT ASUS/Documents/ASRM/Actual
        Dissertation/i5N30SNn30.csv"),"/RC", item.max - item.min + 1, "N", sample.size, "I",
        number.of.items, ".csv")

write.csv(output,"C:/Users/karen/Documents/KT ASUS/Documents/ASRM/Actual
        Dissertation/i5N50normPAreli.csv")

write.csv(population,"C:/Users/karen/Documents/KT ASUS/Documents/ASRM/Actual
        Dissertation/i10N300normPADATA.csv")


output

sd(output)

#skew(population)

#kurtosi(population)
```

APPENDIX C

GENERATING MULTIVARIATE DISTRIBUTIONS FOR CROBACH'S ALPHA-POLYCHORIC ORDINAL ALPHA-MULTILEVEL

GENERATING MULTIVARIATE DISTRIBUTIONS
FOR CROBACH'S ALPHA-POLYCHORIC
ORDINAL ALPHA -MULTILEVEL

Monte Carlo Simulation for MCFA with Cronbch's alpha ad ICCs. For polychoric ordinal alpha use polychori covariance Mtrix and bring it in as with the ACM input below.

```
#----------------Clear All-------------------------#
rm(list=ls())
#--------------SPECIFICATIONS----------------------#
####################################################################

# This code calculates Monte Carlo within and between estimates of reliability (either Cronnbach's alpha or polychoric ordinal alpha) use cor="poly" and bring in at the ACM step
#code uses CFA-derived factor loadings and residual variances.
# Code is for normal distribution with 10 items, n = 300 at level 1 and N = 10 at level-2 and ICCs caclucated at level-2. All output to an excel file to develop 95% CI and SB coefficient to calculate bias.
####################################################################
# Install package MASS
#install.packages("MASS")

# Load package MASS require(MASS)

####################################################################
# User Input

conf <- .95                    # Confidence level (1 – Type I error rate)
reps <- 1000                   # Number of Monte Carlo simulations
set.seed(1842)                 # Set random seed

# Factor loadings and residual variances; used to calculate all reliability estimates wlambda <-
as.matrix(c(.299,.299,.299,.299,.299,.299))                    # Factor loadings within
wtheta <- as.matrix(c(.905,.905,.905,.905,.905,.905))          # Residual variances within
blambda <- as.matrix(c(.137,.137,.160,.160,.183,.183))         # Factor loadings between
btheta <- as.matrix(c(.034,.034,.027,.027,.019,.019))          # Residual variances between
#these can be adjusted as needed
# Input full ACM ordered as lambda(within), theta(within), lambda(between), theta(between)
                    # The ACM can be imported from an external file (as shown), or
                              # inputted directly into the syntax as a matrix
acmMCFA <- as.matrix(read.table("D:\\Traxler\\Monte Carlo CIs\\example.acov"))

####################################################################

#----------          End User Input      ----------------------#

####################################################################

pest <- rbind(wlambda,wtheta,blambda,btheta) # Combine parameter estimates into a single vector nwest <- sum(nrow(wlambda),nrow(wtheta))  # Count number of within-level parameter
```

```
estimates data <- mvrnorm(reps,pest,acmuse,empirical=T) # Generate random draws form joint distribution of
# Parameters
 Nlevel1 = 30
Nlevel2  = 100
ICC = .2
items = 10
rc = 5
inter_cor <- .2

# Compute within-level alphas

# the within-level true score variance estimate as the sum of the within-level
# true-score covariance matrix. Remove the indicator true-score variances
# from the estimated true score variance such that the result equals two times
# each unique coavariance. Divide this result by two to obtain the sum of the
# within-level covariances
wcovs <-  (rowSums(data[,c(1:nrow(wlambda))])^2-rowSums(data[,c(1:nrow(wlambda))]^2))/2

# Find the average within-level covariance
AVGcovWI <- wcovs/((nrow(wlambda)*nrow(wlambda)-nrow(wlambda))/2)

# Compute within-level alphas
WIalpha <- (nrow(wlambda)^2*avgwcov)/(rowSums(data[,c(1:nrow(wlambda))])^2 +
rowSums(data[,c((nrow(wlambda)+1):(nrow(wlambda)+nrow(wtheta)))]))

#Compute the ICCs to calculate the between levels
# Function used to calculate ICC. Single argument requires an lme4 object
ICC_find <- function(model) {
   temp <- VarCorr(model)
   int_var <- (attr(temp[[1]], "stddev")[[1]]) ^ 2
   err_var <- (attr(temp, "sc")) ^ 2
   ICC_temp <- int_var / (int_var + err_var)
   ICC_temp
}

# Recursive function used to reduce highly clustered data set to a desired ICC level
shuffle_ICC <- function(observations, ICC) {
   model <- lmer(tots ~ (1|group), observations)
   ICC_temp <- ICC_find(model)
   if(ICC_temp >= ICC) {
      for(i in 1:j) {
         sample1 <- sample(which(observations$group == i), 1)
         sample2 <- sample(which(observations$group != i), 1)
         group1 <- observations[sample1,"group"]
         group2 <- observations[sample2,"group"]
         observations[sample1, "group"] <- group2
         observations[sample2, "group"] <- group1
      }
      shuffle_ICC(observations, ICC)
   } else if(ICC_temp < (ICC / 2)) {
```

```
      shuffle_ICC(original_mat, ICC)
   } else {return(observations)}

# Print results
# Diagnostics for fun!
y <- sum(apply(observations[,1:items], 2, var))
x <- var(observations$tots)
alpha_c <- (items / (items - 1)) * (1 - (y / x))
model <- lmer(tots ~ (1|group), observations)
ICC_report <- ICC_find(model)
mat_out <- matrix(nrow = 1, ncol = 6, c(j, n, items, rc, alpha_c, ICC_report),
                  dimnames = list(c("Diagnostics"), c("j", "n", "Items","rc", "Alpha", "ICC")))
print(mat_out)
```

APPENDIX D

GENERATING MULTIVARIATE DISTRIBUTIONS FOR PERSON RELIABILITY AT THE SINGLE LEVEL

GENERATING MULTIVARIATE DISTRIBUTIONS FOR PERSON RELIABILITYAT THE SINGLE LEVEL

```
#IRT simulation code #####

IRsim <- function(n_persons = NULL, n_questions = NULL, data_type = NULL, n_cat = NULL,
thresh_var = FALSE, guess_p = NULL, dis_p = 1) {

    n = n_persons # Number of persons
    q = n_questions # Number of question

    person <- rnorm(n, sd = 1) # Person ability range
    item <- rnorm(q, sd = 1) # Item difficulty range
    data <- matrix(nrow = n, ncol = q) # Simulated data frame

    # Dichotomous data - rasch / specify a fixed discrimination parameter. Default set to 1

    if(data_type == "dich") {

        for(i in 1:q) {
            for(j in 1:n) {
                data[j,i] <- rbinom(1, 1, prob = (exp(1) ^ (dis_p * (person[j] - item[i]))) / (1 +
exp(1) ^ (dis_p * (person[j] - item[i]))))
            }
        }
    }
    # Polytonomous data - rasch / default discrimination parameter set to 1 and use RSM

    if(data_type == "poly") {

        item_thresh <- thresh_fun(item, n_cat - 1, thresh_var)

        for(i in 1:n) {
            for(j in 1:q) {
                den <- vector()
                temp_prob <- vector()

                for(z in 1:length(item_thresh[[j]])) {
                    den[z] <- exp(1) ^ sum(dis_p * (person[i] - item_thresh[[j]][1:z]))
                }
                den <- 1 + sum(den)

                for(z in 1:length(item_thresh[[j]])) {
                    temp_prob[z] <- (exp(1) ^ sum(dis_p * (person[i] - item_thresh[[j]][1:z]))) /
den
                }

                temp_prob <- append(1 - sum(temp_prob), temp_prob)
```

```
                data[i,j] <- sample(1:(length(item_thresh[[j]]) + 1), 1, prob = abs(temp_prob))
            }
        }
    }
thresh_fun <- function(item, thresholds, thresh_var) {
    if(length(thresholds) == 1 && thresh_var == FALSE) {
        item_thresh <- lapply(item, function(x) x + seq(from = -2, to = 2, length.out = thresholds))
    } else if(length(thresholds) == 1 && thresh_var == TRUE){
        item_thresh <- lapply(item, function(x) x + seq(from = -2, to = 2, length.out = thresholds)
+ runif(thresholds, min = -.5, max = .5))
    } else if(length(thresholds) != 1 && thresh_var == FALSE){
        item_thresh <- list()
        for(i in 1:length(thresholds)) {
            item_thresh[[i]] <- item[i] + seq(from = -2, to = 2, length.out = thresholds[i])
        }
    } else {
        for(i in 1:length(thresholds)) {
            item_thresh[[i]] <- item[i] + seq(from = -2, to = 2, length.out = thresholds[i]) +
runif(thresholds[i], min = -.5, max = .5)
        }
    }
    item_thresh
}
data_out <- function(data, name, dir = getwd()) {
        setwd(dir)
        write.table(data, file = paste(name, ".csv", sep = ""), sep = ",", row.names = FALSE)
}
 <- (n_persons=30, n_cat=5, n_questions=10, data_type="poly")
#calibrate to winsteps and writeout person reliability (p.rel)
winsteps(u, codes = c(0, 1), noprint = TRUE, ws.path = "C:/Winsteps/", prefix = paste("wstmp",
as.integer(Sys.Date()), sep = ""), peo.mean = 0, item.mean = NULL, scale = 1, peo
data_out<- function(p.rel , name, dir=getwd())
write.table(p.rel, file = paste(name, ".csv", sep = ""), sep = ",", row.names = FALSE)
#Data written out and then looped into Rwinsteps for analysis after changing parameters as
needed above
#output is person reliability for 1000 iterations with specified parameters.
```

APPENDIX E

GENERATING MULTIVARIATE DISTRIBUTIONS FOR PERSON RELIABILITY AT THE MULTILEVEL

GENERATING MULTIVARIATE DISTRIBUTIONS FOR PERSON RELIABILITYAT THE MULTILEVEL

Multilevel data as generated using the Rcode from Appendix C was imported into Winsteps

Example raw data form with level 1 n = 3, level-2 n = 2, 10 items, 5 response choices

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	tots	group
164	4	3	5	4	4	3	5	4	3	4	39	1
833	4	4	4	4	4	4	3	4	3	4	38	1
856	4	5	4	3	4	4	3	4	3	4	38	1
957	3	3	4	4	4	4	5	2	3	4	36	2
994	3	4	3	4	3	3	4	4	4	4	36	2
1400	4	4	4	3	4	4	3	4	3	3	36	2

and the model was specified as:

Model=Multilevel

Rating scale = Multilevel, R5, G, K;

Iteration = 1000

Person reliabity is a part of the summary statistics. It is recorded for each iteration