Ordinal Generalizability Theory Using an Underlying Latent Variable Framework

by

Tavinder K. Ark

B.Sc., McMaster University, 2003
M.Sc., McMaster University, 2005

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Measurement, Evaluation, and Research Methodology)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)
June 2015
© Tavinder K. Ark, 2015
Abstract

This dissertation introduces a method for estimating the variance components required in the use of generalizability theory (GT) with categorical ratings (e.g., ordinal variables). Traditionally, variance components in GT are estimated using statistical techniques that treat ordinal variables as continuous. This may lead to bias in the estimation of variance components and the resulting reliability coefficients (called G-coefficients). This dissertation demonstrates that variance components can be estimated using a structural equation modeling (SEM) technique called covariance structural modeling (CSM) of a polychoric or tetrachoric correlation matrix, which accounts for the metric of ordinal variables. The dissertation provides a proof of concept of this method, which will be called ordinal GT, using real data in the computation of a relative G-coefficient, and a simulation study presenting the relative merits of ordinal to conventional G-coefficients from ordinal data.

The results demonstrate that ordinal GT is viable using CSM of the polychoric matrix of ordinal data. In addition, using a Monte Carlo simulation, the relative G-coefficients when ordinal data are naively treated as continuous are compared to when they are correctly treated as ordinal. The number of response categories, magnitude of the theoretical G-coefficient, and skewness of the item response distributions varied in experimental conditions for: (i) a two-facet crossed G-study design, and (ii) a one-facet partially nested G-study design. The results reveal that when ordinal data were treated as continuous, the empirical G-coefficients were consistently underestimates than their theoretical values. This was true regardless of the number of response categories, magnitude of the theoretical G-coefficient, and skewness. In contrast, the ordinal G-coefficients performed much better in all conditions.
This dissertation shows that using CSM to model the polychoric correlation matrix provides better estimates of variance components in the GT of ordinal variables. It offers researchers a new statistical avenue for computing relative G-coefficients when using ordinal variables.
Preface

Chapter 3 is a stand-alone manuscript to be submitted to a peer-reviewed journal. My contribution was the formulation of the research question, design, computer simulation, data analysis, interpretation of the results and preparation of the manuscript. My supervisor, Dr. Bruno D. Zumbo, helped formulate the question, design the study and prepare the text.
# Table of Contents

Abstract .......................................................................................................................... ii

Preface ............................................................................................................................ iv

Table of Contents .......................................................................................................... v

List of Tables ................................................................................................................ vii

List of Figures ................................................................................................................. viii

List of Symbols and Abbreviations ............................................................................. x

Acknowledgments ........................................................................................................ xii

Dedication ......................................................................................................................... xiv

Chapter 1: Introduction ................................................................................................. 1

Chapter 2: Review of Generalizability Theory and Covariance Structural Modeling ....... 5

  2.1 Reliability Theory ..................................................................................................... 5
  2.2 Classical Test Theory ............................................................................................... 6
  2.3 Ordinal CTT ............................................................................................................ 8
  2.4 Brief Review of Generalizability Theory ................................................................ 9
  2.5 Review of SEM and CSM in Variance Component Estimation ............................. 16
      2.5.1 Pattern of Data Analysis: Q Versus R Methodology .................................... 19
      2.5.2 The Metric of the Observed Variables: CSM Using the Pearson, Tetrachoric and Polychoric Correlations .......................................................... 20
      2.5.3 Estimators ..................................................................................................... 23
      2.5.4 Model Specification ....................................................................................... 26
  2.6 Literature Review of CSM with Continuous and Ordinal Data for GT .................. 28
  2.7 Purpose of the Study ............................................................................................. 29

Chapter 3: Conducting Ordinal Generalizability Theory for Ordinal Response Data ...... 31

  3.1 Introduction ............................................................................................................. 31
      3.1.1 Covariance Structural Modeling for GT Using Continuous Data .................. 33
  3.2 Study 1: Ordinal GT: Covariance Structural Modeling for GT Using Ordinal Data 44
      3.2.1 Methods: Demonstration of Covariance Structural Modeling for GT on a Tetrachoric and Polychoric Matrix Using Ordinal Data ............................. 48
      3.2.2 Proof of Concept: Results ............................................................................. 51
      3.2.3 Discussion ..................................................................................................... 53
3.3 Study 2: A Monte Carlo Simulation Documenting the Effects of Ordinal G-coefficients.............. 55
  3.3.1 Introduction...................................................................................................................... 55
  3.3.2 Methods.......................................................................................................................... 56
  3.3.3 Results............................................................................................................................. 59
  3.3.4 Discussion ...................................................................................................................... 70

Chapter 4: Discussion and Recommendations ............................................................................. 75
  4.1 Novel Contributions to the Field .......................................................................................... 75
  4.2 Implications for Researchers and Assessment Specialists.................................................... 76
  4.3 Limitations and Recommendations...................................................................................... 78
  4.4 Conclusions......................................................................................................................... 79

References.................................................................................................................................... 81

Appendix 1.................................................................................................................................. 86

Appendix 2.................................................................................................................................. 90

Appendix 3.................................................................................................................................. 92

Appendix 4.................................................................................................................................. 93
List of Tables

Table 3.1. Combination of GT and SEM conventional notation for Ordinal GT .................... 36
Table 3.2 Variance Components and relative G-coefficients for EduG's (n = 120) two-facet fully crossed design (prd) ........................................................................................................... 52
Table 3.3 Variance components and relative G-coefficients for EduG’s (n = 100) one-facet nested design (p(i:v)) .......................................................................................................................... 52
Table 3.4. Relative G-coefficients for the theoretical G-coefficient of 0.40 for the crossed design ........................................................................................................................................ 60
Table 3.5. Relative G-coefficients for the theoretical G-coefficient of 0.60 for the crossed design ........................................................................................................................................ 61
Table 3.6. Relative G-coefficients for the theoretical G-coefficient of 0.80 for the crossed design ........................................................................................................................................ 61
Table 3.7. Relative G-coefficients for the theoretical G-coefficient of 0.90 for the crossed design ........................................................................................................................................ 61
Table 3.8. Relative G-coefficients for the theoretical G-coefficient of 0.40 for the nested design ........................................................................................................................................ 62
Table 3.9. Relative G-coefficients for the theoretical G-coefficient of 0.60 for the nested design ........................................................................................................................................ 62
Table 3.10. Relative G-coefficients for the theoretical G-coefficient of 0.80 for the nested design ....................................................................................................................................... 62
Table 3.11. Relative G-coefficients for the theoretical G-coefficient of 0.90 for the nested design ....................................................................................................................................... 63
List of Figures

Figure 2.1. The relationship between y*, y, and the thresholds for a five-point response scale .. 21

Figure 3.1. The path diagram for estimating the variance components in a two-facet fully crossed design (PIO) using SEM. All the one-way paths (structural parameters) are fixed to one so that the variance of the latent variables can be estimated ......................................................... 40

Figure 3.2. Bias (%) of G-coefficients by number of scale points and item responses with no skewness for theoretical reliabilities of 0.90, 0.80, 0.60 and 0.40 for the crossed design.... 63

Figure 3.3. Bias (%) of the ordinal G-coefficients by number of scale points and item responses with no skewness for theoretical reliabilities of 0.90, 0.80, 0.60 and 0.40 for the crossed design .................................................................................................................................................. 64

Figure 3.4. Bias (%) of the G-coefficients by number of scale points and item responses with a negative level of skewness for theoretical reliabilities of 0.90, 0.80, 0.60 and 0.40 for the crossed design .................................................................................................................................................. 65

Figure 3.5. Bias (%) of the ordinal G-coefficients by number of scale points and item responses with a negative level of skewness for theoretical reliabilities of 0.90, 0.80, 0.60 and 0.40 for the crossed design .................................................................................................................................................. 66

Figure 3.6. Bias (%) of the G-coefficients by number of scale points and item responses with no skewness for theoretical reliabilities of 0.90, 0.80, 0.60 and 0.40 for the nested design..... 67

Figure 3.7. Bias (%) of the ordinal G-coefficients by number of scale points and item responses with no skewness for theoretical reliabilities of 0.90, 0.80, 0.60 and 0.40 for the nested design .................................................................................................................................................. 68
Figure 3.8. Bias (%) of the G-coefficients by number of scale points and item responses with a negative level of skewness for theoretical reliabilities of 0.90, 0.80, 0.60 and 0.40 for the nested design........................................................................................................... 69

Figure 3.9. Bias (%) of the ordinal G-coefficients by number of scale points and item responses with a negative skewness for theoretical reliabilities of 0.90, 0.80, 0.60 and 0.40 for the crossed design.................................................................................................................................................. 70
List of Symbols and Abbreviations

Symbols (by alphabetic order)

English Letters
corr (x, x): the correlation between two variables
e: measurement error
E: expected value
S: sample covariance matrix
T: true score
x: observed score
Y: the observed variable(s)
y*: the underlying continuous latent response variable
W: the weight matrix
tr: is the sum of the diagonal elements (trace)
Y_{ijk}: the observed score for the ith person, on the jth variable and kth variable

Greek Letters
ε: the unique error
η: the latent variable
Θ: the diagonal matrix of unique error
θ: the model parameter estimates
Λ: the matrix of structural parameters
λ: the structural parameter estimates (e.g., factor loadings) that connect the observed variables to the latent variables
Eρ^2: the expected reliability of the relative G-coefficient
ρ^{cr}_x: the reliability coefficient of scores on a test
Σ(θ): the model covariance matrix
σ^2: the variance of a random variable (latent or observed)
τ: the thresholds of the underlying continuous latent response variable
Φ: then index of dependability
φ: the covariance matrix of the latent variables

Abbreviations (by alphabetic order)
ADF: asymptotic distribution-free
ANOVA: analysis of variance
CFI: comparative fit index
CTT: classical test theory
CSM: covariance structural modeling
CVM: categorical variable methodology
G-coefficient: generalizability-coefficient
GLS: generalized least squares
GT: generalizability theory
ML: maximum likelihood
MS: mean square method
NT: normal theory
REML: restricted maximum likelihood
RSMEA: root mean square error of approximation
SEM: structural equation modeling
WLS: weighted least squares
WLSM: weighted least squares mean adjusted
WLSMV: weighted least squares mean and variance adjusted
ULS: unweighted least squares
Acknowledgments

I would not only like to thank my supervisor, Dr. Bruno Zumbo, for being the epitome of awesomeness, but for being patient, understanding, caring and a great friend in one of my life’s hardest moments aside from actually writing my thesis. I cannot express my gratitude in words for how lucky I feel to end up having such a wonderful supervisor and mentor. I appreciate the statistical, mathematical and practical application of your supervision. I have come so far in such a short time, statistically, and that can be attributed to your constant, additive belief that I could learn statistics in all its intricacy. I would also like to thank my committee members Dr. Amery Wu and Dr. Sterrett Mercer. Dr. Wu has an ability to articulate complexity in the simplest of ways. I hope one day to be able to achieve the same statistical precision of language as her. I also appreciate the way in which she sees the world outside of statistics. Thank you Dr. Mercer for always reminding me of the practical application of my work; this is essential when developing methods you wish everyday researchers will use. A special thanks to Dr. Daniel W. Cox for helping me get through the last phase of my degree. Working with Dr. Cox has provided me with a new appreciation for writing and attention to each sentence structure. I hope one day to achieve his writing habits. Dan, I really enjoyed working with you and being able to bother you at any time not only for advice, but also to share what I did over the weekend. Thanks for helping me in the next phase of my career as well!

Thank you to HELP for giving me the ‘sweet’ job before I graduated. I enjoy playing in the sand with the nicest and most wonderful people I have met here. To the little man who told me to ‘take it to the house,’ I hope I have done that here.

There are so many wonderful friends close and far to thank. Each individual in his or her own way has provided an important contribution to my thesis even if they do not think so or
know it. To my good friend who said this was a lot of math for such a pretty lady, I appreciate the compliment on all levels. Thank you to Benjamin Shear, Oscar Olvera, Arwa Alkhalaf and Yan Liu for your support as well as statistical and non-statistical conversations. I’ll always ask myself ‘What would Ben do?’ when I think of procrastinating anything. Oscar, my good friend, the number of ways we schemed to find ourselves space will always make me laugh. Arwa, thank you for the warm hugs and tender smile whenever I needed them. Thank you Yan for making me feel right at home in the lab and teaching me great things come to nice people.

A special thanks to the ‘B.C’ family. I’ve appreciated our relationship from laughing out of control at Dasio, the correction of my Punjabi, late night photo shoots, conversations, studying, randomness and hospital visits. Although you’ve proven to me I’m no model behind the camera when posing, you’ve shown me the comfort of your home. Each one of you has provided me with an endless list of fond memories that made my life fuller during my studies at UBC.

On a personal note, I’d like to thank my dad for making everything possible and letting me sweat the small stuff, my brother, for his steadfast, unquestionable support and always ensuring I tried every dessert at Whole Foods. None of this would ever be possible without the unwavering confidence of my mom. It is hard to thank someone so monumental in my life as her. I’d like to dedicate this thesis to her. My mom has taught me many things she will never realize. Of the things I am most proud of is that she has taught me what waheguru is in the face of adversity, to never feel sorry for yourself because the struggle you are up against is what makes you what you are. Just laugh and let it go. I am going to close my acknowledgments with all of the things I forgot to say racing through my mind until we play cards again.
To my mom
Chapter 1: Introduction

The use of generalizability theory (GT) to conceptualize and estimate “dependability” has steadily increased in educational and psychological research. GT is a statistical technique that expands upon classical test theory (CTT) in that it separates the single undifferentiated error term into multiple sources of error in one analysis. It also focuses on the generalizability of observed scores to a universe of observations. As Cronbach, Rajaratnam and Gleser (1963) note, GT was introduced as a liberalization and reinterpretation of reliability theory. Over the next fifty years, some proponents of GT created terminology that distinguishes it from CTT (e.g., Brennan, 2001). Although these alternative terms address the need for precision, they may confuse users of GT because they are “difficult to define in the abstract” (Feldt & Brennan, 1989, p. 128). Like Cronbach et al. (1963), modern practitioners and assessment specialists appear to treat GT as providing new formulas for estimating “reliability” (Crocker & Algina, 1986, p. 158). Therefore, throughout this dissertation, the terms “G-coefficients” and “coefficients of dependability” will be used interchangeably with “reliability coefficients,” and the generic term “reliability” with “generalizability.” Doing this also helps adapt and extend concepts from nonlinear CTT for estimating ordinal reliability coefficients (Zumbo, Gadermann, & Zeisser, 2007) to GT, which is a primary focus of this dissertation.

GT allows for the estimation of variance components based on the assessment (sometimes called the measurement) design. These variance components are essential to the calculation of generalizability coefficients (G-coefficients), which are meant to aid in evaluating the generalizability (dependability) of a measurement. The primary way in which variance components are estimated in practice is by using the mean squares (MS) method via a factorial or
repeated measures analysis of variance (ANOVA), although other estimation methods and model parameterizations may be used.

Variance components can be obtained using conventional linear models with random factors (e.g., Searle, Casella, & McCulloch, 2006), or by using structural equation modeling (SEM) of a covariance matrix (e.g., Bock & Bargmann, 1966). These two approaches draw on various estimation methods, such as maximum likelihood (ML), robust variants of ML, variations on weighted least squares (WLS), and Bayesian approaches. Both conventional linear models with random factors and SEM assume that the dependent variables are continuous (Finney & DiStefano, 2006; Rhemtulla, Brosseau-Liard, & Savalei, 2012; Searle et al., 2006).

This state of affairs in variance component estimation is alarming, considering that most of the data analyzed in GT is (ordered) categorical. Constructs are measured on an ordered rating scale, such as a binary (e.g., incorrect/correct or present/absent) or polytomous response scale (e.g., strongly disagree, neither agree nor disagree, strongly disagree). When employing conventional linear models with random factors (e.g., Brennan, 2001) or SEM (e.g., Marcoulides, 1996) to compute variance components in GT, the metric of the ordinal variables is ignored and the data are treated as if it was continuous in the analysis. In general, this may lead to spurious hypothesis testing, biased estimations of statistics, incorrect inferences and the misrepresentation of the relationship between the ordinal variables and its relationship to the latent variables (Holgado-Tello, Chacón-Moscoso, Barbero-García, & Vila-Abad, 2010; Jaeger, 2008; Muthén, L., 1983). More specifically, researchers have shown that ignoring the ordinal metric of the data can lead to biased estimates in the coefficient alpha index of reliability (Gadermann, Guhn, & Zumbo, 2012; Zumbo et al., 2007). This suggests that when dealing with ordinal variables, statistical analyses that assume the data to be continuous should be avoided in
the estimation of variance components and the subsequent computation of the reliability indices in GT.

A way to model the metric of ordinal data is using an SEM technique called covariance structural modeling (CSM). Bock (1960) and Bock and Bargmann (1966) introduced CSM in the estimation of variance components for continuous data, which is akin to a random effects ANOVA. Following in their footsteps, Marcoulides (1996) and others demonstrated that CSM could be used to estimate variance components for GT. However, these researchers either dealt with continuous variables or followed the tradition in GT treating ordinal data as continuous. L. Muthén (1983) was the first to correctly model the ordinal metric of binary response data in the computation of variance components for GT in her dissertation. She accomplished this by modeling the tetrachoric correlation matrix for the binary item responses using CSM. Her results, along with those of Zumbo et al. (2007), suggest that treating ordinal data as continuous can lead to substantially inaccurate and deflated reliability estimates. To the author’s knowledge, L. Muthén (1983) is the only scholar to date to explore the computation of variance components for GT by correctly treating binary data as ordinal. The goal of this dissertation is to build on L. Muthén’s work by applying this approach to polytomous data and more complex assessment designs, such as those involving nested components.

The purpose of this dissertation is to provide proof of concept that CSM of the polychoric correlation matrix can be used to estimate the variance components needed to make a relative (e.g., rank-order) decision in GT with ordinal response data and complex assessment designs. It also aims to compare the relative G-coefficients when ordinal data are treated as continuous (GT) and when it is treated as categorical (ordinal GT) using CSM. It compares both of these coefficients against their theoretical values. It examines two designs: a two-facet fully crossed
design and a one-facet partially nested design. Estimates of the relative G-coefficients are compared when varying the number of response categories, the magnitude of the theoretical G-coefficients, and the skewness of the item response distributions. The results will have important practical implications for researchers in deciding which methods are most effective when dealing with ordinal data in GT.

This dissertation is organized as follows. Chapter 2 provides a brief overview of reliability theory, emphasizing the pitfalls of CTT and the benefits of GT. This is followed by a summary of CSM and tetrachoric and polychoric correlations. Chapter 3 consists of two studies. Study 1 is a proof of concept of conducting ordinal GT using real data sets consisting of binary or polytomous item responses. Study 2 compares the G-coefficients and ordinal G-coefficients of ordinal data to theoretical G-coefficients using a Monte Carlo simulation. Chapter 4 discusses the major findings of this dissertation, along with its novel contributions, limitations, real-world implications for researchers, and future directions.
Chapter 2: Review of Generalizability Theory and Covariance Structural Modeling

2.1 Reliability Theory

Reliability is used to quantify the amount of measurement error and the consistency of a set of test scores to measure the same thing across repeated attempts (Allen & Yen, 1979). In practical terms, reliability is the degree to which individuals’ scores remain consistent over repeated observations of the same test or alternative test forms. To some extent, all tests are unreliable because an examinee’s score on a test of items only represents a limited sample of their behaviour. As a result, test scores obtained under these conditions are fallible and subject to measurement error. The extent of this unreliability is a cause of concern for researchers and test developers.

There are two sources of measurement error: systematic and random. Factors that bias individuals’ observed scores are referred to as systematic sources of error or simply as systematic errors (Hoyt & Melby, 1999). Formally, systematic errors are defined as errors “which consistently affect an individual’s score because of some particular characteristics of the person or the test that has nothing to do with the construct being measured” (Crocker & Algina, 1986, p. 105). Systematic errors persist across repeated testing with the same instrument and influence an individual’s score in a consistent manner. Examples include item and rater biases. However, it is easy to forget that not all systematic variance is due to the object being measured (e.g., the test taker) or the construct of interest.

By contrast, random errors affect an individual’s test score because of “purely chance happenings” (Crocker & Algina, 1986, p. 106), and include guessing, distractions and fluctuations in mental state. Systematic errors do not result in inconsistent measurement,
although they may cause inaccurate representation and interpretation of test scores. However, random errors reduce the consistency and usefulness of test score interpretation. Both systematic and random errors together are referred to as measurement error in CTT, but in GT, systematic errors are often called sources of error or sources of variation. This dissertation will refer to systematic errors as either systematic errors or sources of error to prevent confusion.

Two notable statistical theories for conceptualizing and quantifying reliability are CTT and GT. Both offer a model for estimating the amount of measurement error in the observed scores for a particular group of individuals.

2.2 Classical Test Theory

CTT is most commonly employed conceptualization of reliability. CTT, assumes that an individual’s observed score \(x\) on a measure or a test is an additive combination of his/her true score \(T\) and measurement error \(e\). The observed score in CTT theory can be written as:

\[x = T + e.\]

The true score \(T\) reflects an individual’s score on a test/measure, or the average score across a large number of replicated (strictly parallel) measurements of a particular construct/attribute. The measurement error captures all sources of random and systematic error in the observed score (Hoyt & Melby, 1999). The true score is the expected value of the observed score. Thus, in CTT, \(E[x] = T\) and hence \(E[e] = 0\), where \(E[]\) denotes the expected value. As a consequence of these statements about the expected values, \(\text{Corr}(T, e) = 0\) and \(\text{Corr}(e_1, e_2) = 0\), where \(\text{Corr}()\) represents the correlation between two random variables, \(e_1\) is the error arising from one test, and \(e_2\) is the error arising from a parallel test (Allen & Yen, 1979). Simply put, the true and error scores are uncorrelated. This means that examinees with low scores do not consistently have more positive or negative errors of measurement than those with high scores. Furthermore, the error scores on
two different measures are uncorrelated, although an individual may have a negative error score on one test, this does not mean that he/she is more likely to have a negative error score on another. The measurement error is uncorrelated with an individual’s true score as well as with other measurement errors.

In general, reliability coefficients range from zero to one, where one indicates that the observed scores are free from measurement error and zero indicates complete unreliability. Reliability can be quantified and interpreted in various ways in CTT. Formally, the reliability of a set of test scores, \( x \), can be written as:

\[
\text{rel}(x) = \rho_{xT}^2 = \frac{\sigma_T^2}{\sigma_x^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_e^2},
\]

where \( \rho_{xT}^2 \) is referred to as the reliability coefficient of scores on a test and the true, observed, and error score variances are denoted as \( \sigma_T^2 \), \( \sigma_x^2 \), and \( \sigma_e^2 \), respectively. The reliability coefficient can generally be interpreted as the ratio of the true to the observed score variance. These CTT definitions and ratios allow researchers to quantify and estimate the reliability of a set of test scores for a particular group of individuals in various ways.

Different reliability coefficients can be estimated depending on which source of measurement error is to be quantified. The most commonly estimated reliability coefficients are a) the temporal instability of the observed score (test-retest), b) heterogeneity among items (internal consistency), c) consistency across raters (inter-rater), and d) heterogeneity among different forms of the same test (parallel test forms). Researchers often report multiple reliability coefficients when multiple sources of measurement error may influence the observed score (Allen & Yen, 1979).
2.3 Ordinal CTT

Conventional CTT involves reliability coefficients that assume that the observed item responses are continuous. For example, the widely used coefficient alpha (internal consistency) is, in essence, a function of variances and covariances that presupposes the continuity of the variables (items in this case). This issue has been, until recently, largely ignored in the psychometric literature. Zumbo et al. (2007) developed ordinal coefficient alpha (and theta) as a new method to compute coefficient alpha when item responses are ordinal. Exploiting the connection between factor analysis and reliability theory, the researchers obtained ordinal coefficient alpha using factor analysis of the polychoric correlation matrix. In this technique, the metric of the ordinal data are considered via an underlying latent variable approach (the polychoric correlation).

To investigate the benefits of their new technique, Zumbo et al. (2007) conducted a simulation study comparing the theoretical reliability of conventional coefficient alpha and ordinal coefficient alpha using rating scale (Likert) data. They varied the number of response categories, theoretical reliability, and skewness of the item distributions. The results indicated that regardless of the magnitude of the theoretical reliability and number of response categories, conventional coefficient alpha consistently underrepresented the true reliability, especially when the item response distributions were skewed. This was not the case for ordinal alpha, which consistently produced an unbiased estimate of the theoretical reliability. These findings suggest that ordinal alpha should be computed when using ordinal item responses for CTT. It is therefore reasonable to expect that G-coefficients may fare no differently when the ordinal metric is ignored in the analysis. Having described the fundamentals of CTT, we will now briefly review GT.
2.4 **Brief Review of Generalizability Theory**

Although CTT provides an elegant framework for measuring the impact of measurement error, it is rather simplistic resulting in an inadequate estimation of the total proportion of error in test scores for a given population. CTT is criticized for the ambiguity of its concept of true score variance and the identification of the various sources of measurement error (Hoyt & Melby, 1999). Both of these criticisms limit the interpretation of reliability coefficients.

A measure may have multiple “true” scores depending on its intended application, and may thus have multiple reliability coefficients. For example, one researcher may be interested in test-retest reliability, where the true score generalizes over time, while another may wish to examine inter-rater reliability, where the true score generalizes among raters. What is considered sources of measurement error vary depending on what the researcher is investigating and therefore, it makes little sense to label all variance as either systematic or random.

In CTT, the measurement error term applies to both random and systematic errors and is thus considered undifferentiated. Furthermore, CTT cannot explore the interaction effects among the sources of error. This latter issue is important because it means that CTT cannot investigate the influence of the various sources of error on the observed scores and may generate a gross reliability estimate. Within the CTT framework, researchers may be confused on how to calculate the overall impact of measurement error because they can use more than one reliability coefficient. GT overcomes these shortcomings and extends reliability theory by estimating both systematic and random sources of measurement error simultaneously in one analysis.

GT provides a conceptual framework and statistical theory for evaluating the dependability of a set of observed scores for a given measure in different universes (Cronbach et al., 1963; Hoyt & Melby, 1999; Shavelson, Webb, & Rowley, 1989). In GT, the observed scores
obtained from a measure are treated as a sample from a universe of admissible observations, which consists of all possible observations that could be considered interchangeable (or equivalent) to those actually obtained. The universe of admissible observations is bounded by all possible combinations of the levels of the facets, which are defined as the measurement conditions (e.g., raters, items, occasions). The universe of generalization is the levels of the facets in the measurement design over which one wishes to generalize. GT describes the accuracy of generalization made from observed scores to universe scores of individuals. A universe score is the expected value of the observed score over all observations in the universe of generalization and can be considered analogous to the true score in CTT (Shavelson & Webb, 1991). GT describes how accurately the observed scores generalize to the universe scores of individuals. When measuring a construct, different investigators may be interested in different universes of generalization and universe scores. Thus, reliability theory is reinterpreted as a theory on the accuracy with which researchers can generalize from one observation to a universe of observations (Cronbach et al., 1963).

To evaluate the generalizability of a set of observed scores, a generalizability (G) study is conducted to estimate the variance due to each source of error and the universe score, which is referred to as the object of measurement. The object of measurement is the facet measuring the construct of interest. People are normally the object of measurement because researchers are interested in designing assessments that systematically differentiate between individuals on an underlying construct in a systematic manner (Brennan, 2001; Hoyt & Melby, 1999). All other facets that cause variation in the observed scores, other than the object of measurement, are
considered sources of error in GT.\(^1\) The universe of generalization identifies the likely sources of error in a measurement design. The most common of these are raters, items, cases, time, domains, and all of the interactions with the object of measurement and among themselves. The random error term consists of both random error and the highest-order interaction term between the facets. The basic approach to a G-study is to decompose the total variance in the observed score into the object of measurement (universe score), source(s) of error, and random error (Shavelson et al., 1989). Using the information from the G-study, a decision study (D-study) can be conducted to help re-design the measurement tool to minimize the impact of the sources of error on the observed score.

GT evaluates the effect of the various sources of error in a given measurement design by estimating the generalizability coefficients, often referred to as the G-coefficients. A G-coefficient is a ratio of the universe score variance to the total variance; it estimates the proportion of the total variance that is attributable to the universe score (Hoyt & Melby, 1999). G-coefficients are an index of the extent to which a set of observed scores on a given measure would generalize to other, functionally equivalent, measurement conditions (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Hoyt & Melby, 1999).

Researchers must consider a few important steps and design concepts before estimating variance components and G-coefficients. The first step in conducting a G-study is to explicitly identify the sources of error that may affect the observed score. Next, one should determine a) if a relative or absolute decision will be made from the observed scores, b) if the factors are fixed or random, c) if any of the factors are crossed or nested in the measurement design, and d) the

---

\(^1\) It is important to note that “facet” is a unique term in GT, and is equivalent to a factor in linear models and a latent variable (often called a factor) in SEM or CSM. “Facet” will be reserved for explaining the design of a G-study, while “factor” will be used everywhere else to prevent confusion.
estimation method to be used. Identifying these aspects will determine what factors will be
included in the denominator of the G-coefficient and whether all sources of error can be
partitioned from the random error term.

The two most common estimated G-coefficients are the generalizability coefficient and
index of dependability. Each one of these coefficients reflects the type of decision that will be
made from a set of observed scores (Brennan, 2001). The generalizability coefficient (denoted
as $E\rho^2$) is estimated for a relative decision and is referred to as the relative G-coefficient. A
relative decision compares the differences among the observed scores of individuals (e.g.,
individuals are ordered by rank). In estimating the relative G-coefficient, all the interaction terms
with the object of measurement contribute to the rank ordering of individuals and are considered
sources of error. The relative G-coefficient is calculated as follows:

$$E\rho^2 = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_{relative}^2},$$

where $E\rho^2$ is the expected reliability of the relative G-coefficient, $\sigma_r^2$ is the variance due to the
object of measurement and $\sigma_{relative}^2$ is a term that contains all the interactions with the object of
measurement and random errors (Brennan, 2001).

The index of dependability (denoted as $\Phi$) is estimated for an absolute decision and is
referred to as the absolute G-coefficient (Brennan, 2001). An absolute decision is when the
observed scores are interpreted without reference to other individuals’ performance (e.g., passing
or failing an exam based on a cutoff). In estimating the absolute G-coefficient, all the variance
components and interactions are considered error because each factor contributes a constant
effect on the observed score. For example, the stringency of a rater will influence an individual’s
observed score, and hence the absolute decision about him or her. Consequently, the error due to
the rater is included in the denominator if it is not fixed. The absolute G-coefficient is calculated as follows:

$$\Phi = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_{absolute}^2}$$

where $\Phi$ is the expected reliability of the absolute G-coefficient, $\sigma_r^2$ is the variance due to the object of measurement, and $\sigma_{absolute}^2$ contains all the sources of error (including random error) and their interactions (Brennan, 2001). In summary, an absolute decision involves some form of a cut-off score to which each individual is compared, while a relative decision orders people by rank. The latter type of decision has fewer sources of error.

There are two ways to determine whether a factor is fixed or random. First, one must consider how a set of observed scores is to be generalized. If it is to be generalized to other universes of observations (e.g., situations, behaviours or abilities), then the factor is random. However, if the scores are not to be generalized beyond the context in which they are obtained, then the factor is fixed. Secondly, one must decide whether each level within a factor is exchangeable with all of the other levels within that factor. A factor is considered random when the levels are interchangeable. For instance, a researcher may select any item from a pool of items to be included on a test. In this case, the item is random because it is interchangeable with any other item from the same pool. Therefore, a factor is random if it is exchangeable and generalizable. On the other hand, a factor is fixed when its levels are exhausted or cannot be exchanged. For instance, if the items on a test cannot be replaced by any others and the researcher does not wish to generalize beyond these specific items, then they would be considered a fixed factor (Shavelson & Webb, 1991).
The variance due to the sources of error changes depending on whether the factors are treated as random or fixed. Computationally, the variance due to a random factor as a main effect is included in the denominator for an absolute decision, but it is not for a relative decision. When a factor is fixed, new variance components must be computed for all of the terms that interact with it. Although the interaction terms may end up in the denominator depending on the decision, the main effects due to the fixed factor are not included in the denominator of an absolute G-coefficient.

Another issue to consider in GT is whether the research design is crossed or nested. A facet (e.g., A) is nested within another facet (e.g., B) if the multiple levels of A are associated with only one level of B. On the other hand, a design is crossed when all levels of A occur in all levels of B. For instance, in a crossed design, all raters will observe all examinees for each item, whereas in a nested one, raters may observe only a subset of examinees. As a result, “rater” is nested within “examinee.” Although a nested design is cost effective, it can pose unique challenges in computing variance components because it is impossible to estimate, within a conventional ANOVA framework, how much variance is due to the nested facet. For this reason, a crossed design is preferred (Brennan, 2001; Shavelson & Webb, 1991).

Variance components are both the building blocks and the Achilles’ heel of GT. While there are many ways to compute them, the most common, due to its ease of use, is the mean squares (MS) method, which uses a factorial or repeated measures ANOVA. In this method, variance components are estimated by equating the observed MS from an ANOVA to the expected values. This is done by solving a set of linear equations for each identified source of error, including the main effects, interaction terms and residual error term. The greatest problem with using the MS method is that it can result in unstable and negative variance component
estimations (Searle et al., 2006). Cronbach et al. (1963) suggest setting the negative variance components to zero and using the value of zero in the MS equations of other variance component calculations. Brennan (1983) also recommends setting the negative variance component(s) to zero, but only after using the negative value in the MS equation to calculate the other variance components. Although Brennan’s (1983) recommendation leads to unbiased estimates (Shavelson et al., 1989), neither solution explains why negative variance components occur. Indeed, such negative values suggest that the sample size may be too small or that the model is poorly specified. These problems may introduce biases and inaccurate estimates of the G-coefficients.

More powerful methods than the MS exist for estimating variance components, such as the use of a linear model with random factors (e.g., Searle, Casella, & McCulloch, 2006), or structural equation modeling (SEM) of a covariance matrix (e.g., Bock & Bargmann, 1966). Either class of model parameterization (conventional linear models with random factors or SEM) can draw on various estimation techniques such as maximum likelihood (ML) and robust variants thereof, variations on weighted least squares (WLS), and Bayesian methods. It is important to note that both linear models and SEM can be used to model the metric of ordinal data depending on the choice of estimator. However, the conventional versions of these techniques assume that the dependent variables are continuous (Finney & DiStefano, 2006; Rhemtulla et al., 2012; Searle et al., 2006).

Now that we have reviewed the fundamentals of CTT, GT, and the conventional approaches to variance components estimation, let us now turn to the estimation of variance components using CSM.
2.5 Review of SEM and CSM in Variance Component Estimation

SEM is a statistical method used to specify, test and estimate linear models of the relationships among variables. The relationships under study are represented by a series of structural (e.g., regression) equations and can be shown pictorially. The entire system of variables can be tested statistically in one analysis to determine the extent to which the model is consistent with the data. If the model fits the data, then the postulated relationships among the variables are plausible; if not, however, the model is rejected (Bollen, 1989).

A full model in SEM consists of a system of structural equations that contain random variables and structural parameters. The random variables (often referred to as factors) may be latent and/or observed, and the measurement scale may be continuous or discrete (e.g., nominal or ordinal). The structural parameters (often referred to as estimated or model parameters, or factor loadings) are constants that describe the relationship between the variables in the model. The system of structural equations has two subsystems: the measurement model and the structural model. The latter includes the equations that describe the linear relationship between the latent variables (η) and/or observed variables, while the former includes those that describe the linear relationship between the observed (Y) and the latent variables (Bollen, 1989).

Let us use the common factor model to explicate some fundamental concepts and notations in SEM. The set of structural equations representing this model may be summarized by a single equation as follows:

\[ Y = \mu + \Lambda \eta + \varepsilon, \]

where p (observed variables, (Y)) is modeled as a linear combination of m (latent variables, (η)) and unique error (ε). The mean of the population of participants (μ) may be included in the model. The \( \Lambda \) consists of a matrix of estimated structural parameters describing the effects of
the latent variables ($\eta$) on the observed variables ($Y$). The $\varepsilon$ consists of random measurement error and item-specific error in each of the $p$ observed variables. Together, these two types of error form the unique error, $\varepsilon$, of the observed variables and are considered error in measuring the latent variables. For this reason, this term is simply referred to as unique error.

The covariance matrix, $\Sigma(\theta)$, of the $p$ observed variables ($Y$), written as a function of the unknown model parameters, $\theta$, can be decomposed as follows:

$$\Sigma(\theta) = \Lambda \varphi \Lambda' + \Theta_\varepsilon,$$

where $\Sigma(\theta)$ is a $p \times p$ model covariance matrix based on the estimated model parameters $\theta$ of $p$ observed variables, $\Lambda$ is a $p \times m$ matrix of the structural parameters ($\lambda$) connecting the observed ($Y$) to the latent variables ($\eta$), $\phi$ is an $m \times m$ square and symmetrical covariance matrix of the latent variable ($\eta$), and $\Theta_\varepsilon$ is a diagonal matrix of the unique error ($\varepsilon$).

Conceptually, SEM decomposes the error in the observed variables into common and unique variance. Common variance is the variance in the observed variables due to the latent variable(s). Unique variance consists of random error variance and specific variance. Specific variance is not explained by the latent variable(s), but is attributable to the observed variable itself; it cannot be disentangled from measurement error unless a measure of reliability is available (Jöreskog, Andersen, Laake, Cox, & Schweder, 1981).

SEM models the sample covariance matrix between the observed variables to measure their linear association. A covariance matrix is a square and symmetrical matrix with the variance among the observed variables on the main diagonal and the covariances on the off-diagonal (a correlation matrix is a standardized covariance matrix with the original units of measurement removed). The model covariance matrix is a function of the structural parameters.
Researchers can test the plausibility of the model covariance matrix by statistically determining the extent to which the model is consistent with the data. The equation can be summarized as follows:

\[
Data(S) = Model(\Sigma(\theta)) + \text{Residual}.
\]

The data represents the covariance among the observed variables (sample covariance matrix); the model is the model covariance matrix, \((\Sigma(\theta))\), written as a function of the parameters \((\theta)\); and the residual is a matrix that captures the difference between the sample \((S)\) covariance matrix and the model covariance matrix. How well the model fits the data can be calculated using exact fit indices, such as the chi-square goodness of fit, or approximate fit indices, such as the root mean square error of approximation (RMSEA) and the comparative fit index (CFI). The model is assessed based on whether the model covariance matrix—a function of the parameters—consistently reproduces the sample covariance matrix using model fit indices. If the parameters and the model covariance matrix are correct, then the parameters will reproduce the sample covariance matrix and the off-diagonal elements of the residual matrix will be approximately zero.

SEM utilizes a number of statistical procedures to investigate the relationship between the variables in a model. CSM, as a type of SEM, analyzes the covariance among the multinormal observed variables to estimate the variance due to the latent variables (Bock & Bargmann, 1966; Bock, 1960; Bollen, 1989). However, unlike other SEM techniques, CSM does not include the mean of the observed and latent variables. The purpose of CSM is to estimate the variance due to the latent variables and unique error, which is achieved by parameterizing the model in a specific manner.
To accurately estimate the variance components of the latent variables, the data analyzed and model must be specified in a particular manner when using CSM. First, one must determine how the observed variables are represented in the sample covariance matrix by considering the metric and pattern of relationships among the observed variables. Next, one should specify the estimation method that is appropriate for the distribution and metric of the observed variables. Finally, one should specify the model in a way to estimate the variance components of the latent variables. All three of these specifications are vital for producing the most accurate and unbiased variance components estimation. Given the purpose of this dissertation, each one of these requirements will be discussed in further detail.

2.5.1 Pattern of Data Analysis: Q Versus R Methodology

The goal of SEM and CSM is to understand the covariance among the observed variables and decompose it into common and unique variance. Two patterns of relationships between the observed variables can be modeled: the Q- and R-methods (McKeown & Thomas, 1988). A Q-method analyzes the covariance between individuals or groups (e.g., person 1 correlated with person 2) across a sample of observed variables. For example, if a researcher wishes to understand the similarities among individuals or groups, then the variables, not the individuals, are the sample. An R-method, by contrast, examines the covariance between the observed variables (e.g., item 1 correlated with item 2) across a sample of individuals. For example, a researcher would use this method to reveal patterns of variation in individuals’ characteristics. The variance and covariance among items—the R-method—is the most common way of analyzing the relationships among observed variables in psychology, education and behaviour research (McKeown & Thomas, 1988). The type of method—whether Q or R—is a very
important specification because it dictates which variance components can be estimated for a
given measurement design using CSM for GT.

2.5.2 The Metric of the Observed Variables: CSM Using the Pearson, Tetrachoric and
Polychoric Correlations

Categorical data are often used in psychology, education and the behavioural sciences
(Jaeger, 2008), yet most statistical analyses, including GT, treat it as continuous. When observed
variable responses are viewed in this manner, the covariance matrix is estimated using a series of
Pearson correlations between the observed variables in the data set (e.g., Pearson correlations
between items 1 and 2, items 1 and 3, etc.). One of the major assumptions in using Pearson
correlations between two observed variables is that the item responses are continuous. The
resulting covariance matrix is symmetrical, comprised of the covariances between the variables
on the off diagonals and the variances of the variables on the diagonal. In this context, a Pearson
correlation matrix is a standardized covariance matrix.

When ordinal variables are treated as categorical, it is assumed that for each variable, y,
there is an underlying continuous latent response variable called $y^*$, which is normally distributed
with a specified mean and variance (Bollen, 1989; Jöreskog, 1994). The points that divide the
continuous latent response variable into discrete categories (c) are called thresholds ($\tau_c$). The
total number of thresholds will equal the number of categories minus one (c-1). The thresholds
on the underlying latent response variable may be divided at unequal intervals and the underlying
distribution may not be normal—although a normal distribution is commonly assumed for
simplicity and ease of computation.

As stated earlier, the thresholds are the values that divide the underlying continuous latent
response variable ($y^*$) into categories to obtain the ordinal variable (y). For example, the
observed ordinal response for item 1 \( (y_1) \) with five response categories is defined by the underlying variable \( y^* \), such that:

\[
y_1 = \begin{cases} 
1 & \text{if } y^* \leq \tau_1 \\
2 & \text{if } \tau_1 < y^* \leq \tau_2 \\
3 & \text{if } \tau_2 < y^* \leq \tau_3 \\
4 & \text{if } \tau_3 < y^* \leq \tau_4 \\
5 & \text{if } y^* > \tau_4
\end{cases}
\]

A participant’s score on the latent response variable \( (y^*) \) is much more precise than on the five-point response scale (see Figure 2.1), which can only be discrete values from one to five. The ordinal response scale can thus be thought of as providing an approximation of the score on the underlying continuous latent response variable \( (y^*) \).

![Figure 2.1. The relationship between \( y^* \), \( y \), and the thresholds for a five-point response scale](image)

When variable responses are ordinal and treated as such, the sample covariance matrix in SEM is a tetrachoric or polychoric correlation matrix. A tetrachoric correlation represents the relationship between two ordered response categories (e.g., yes/no or 0/1), whereas a polychoric correlation represents the relationship between polytomous data (i.e., three or more response categories). Tetrachoric and polychoric correlations are estimated from the bivariate distributions.
of each threshold between two latent response variables, $y^*_1$ and $y^*_2$, of the corresponding ordinal variables, $y_1$ and $y_2$. Simply put, tetrachoric and polychoric correlations are among the underlying latent variables of the ordinal variables. Using these thresholds to estimate the sample covariance matrix is much more precise than using either the actual ordinal response categories (e.g., 1, 2, 3, 4 or 5 on a five-point response scale). The thresholds and correlations between the underlying latent response variables are estimated either separately or simultaneously.

One caveat to consider when using tetrachoric and polychoric correlations is that the model assumes that the latent response variable is continuous. As a result, this type of correlation may not be appropriate if it is truly discrete (Uebersax, 2011). Therefore, tetrachoric and polychoric correlations presuppose that the latent response variable is continuous and linearly related to the dependent variable (Holgado–Tello et al., 2010).

There are various theoretical and empirical reasons why researchers and statisticians should avoid using the Pearson correlation to examine the degree of association between ordinal variables. Theoretically, ordinal variables imply ordinal scales and therefore require statistical analyses that can accommodate the ordinal metric. Not only does the Pearson correlation assume continuous response categories, but it also typically reduces the variability and covariance of ordinal data (Holgado–Tello et al., 2010). This may cause biased parameter estimates when using SEM techniques that do not take the metric of the ordinal variables into account.

Empirically, extensive research shows that when the metric of ordinal data are ignored, the parameter estimates, hypothesis testing, and variance components are incorrectly calculated and impossible to interpret (Holgado–Tello et al., 2010; Jaeger, 2008; Muthén, 1983; Rhemtulla et al., 2012; Wirth & Edwards, 2007). Many studies confirm that at least a five-point response scale is required to use Pearson correlations with ordinal data, though this may be untrue when
the distribution of the item responses is extremely skewed (Gadermann et al., 2012; Rhemtulla et al., 2012; Zumbo et al., 2007). This body of evidence suggests that the metric of categorical variables should not be ignored, especially when the data are not normally distributed or the number of response categories is less than five.

### 2.5.3 Estimators

Estimators are used to compute a single-point or range (interval) estimation of plausible population parameters from sample statistics. Two broad classes of estimators relevant to SEM are the normal-theory (NT) and asymptotic distribution-free (ADF) methods, which employ categorical variable methodology (CVM; Brown, 1984; Finney & DiStefano, 2006; Muthén & Kaplan, 1985; Muthén, 1984). The former is designed for continuous multivariate normal data, while the latter is intended for ordinal data.

NT estimators assume that data are continuous and from a multivariate normal population. The most common of these are ML and generalized least squares (GLS). Both estimators involve an iterative process where the discrepancy (fit) function is minimized between the estimated and sample covariance matrix. That is, the final estimated parameters reduce the difference between the sample \( S \), and the model covariance matrix calculated from the model’s parameters, \( \Sigma(\theta) \). The fit function for both ML and GLS can be written as follows:

\[
F = \frac{1}{2} \text{tr}([S - \Sigma(\theta)]W^{-1})^2,
\]

where \( \text{tr} \) (trace) is the sum of the diagonal elements and \([S - \Sigma(\theta)]\) is the residuals, which are the differences between the elements in the sample and the model covariance matrix (Finney & DiStefano, 2006). These residuals are weighted by an inverted weight matrix, \( W^{-1} \), although this

---

2 It should be noted that the standard practice in the literature of describing ‘bell-shaped symmetric ordinal data’ as ‘normal’ is followed even though ordinal data, by definition, is not normal.
matrix varies between ML and GLS. In ML, W is a function of the estimated model covariance matrix, while in GLS, it is a function of the sample covariance matrix and the second-order product moments of the observed variables (e.g., the sample). Variants of ML, such as restricted maximum Likelihood (REML), exist to account for non-normal distributions. REML, in particular, accounts for fixed effects and can produce unbiased estimates of variance and covariance parameters. However, these estimators may be less reliable than their ADF estimators, especially when the data are overly skewed and/or ordinal.

On the other hand, ADF estimators do not assume multivariate normality in the population and may be used on either continuous or categorical data. Since most categorical data are not normally distributed, ADF estimators are often used. The most common ADF estimator is weighted least squares (WLS). As previously stated, ignoring the metric of ordinal data leads to biased parameter estimates. To counteract this, B. Muthén (1984) developed CVM to include the ordinal metric. This theory utilizes the correct weight matrix (i.e., an inverted tetrachoric or polychoric matrix) when using ADF estimators.

The fit function for WLS is as follows when the metric of the ordinal data are ignored and the data non-normal:

\[ F_{WLS} = [s - \sigma(\theta)] W^{-1} [s - \sigma(\theta)], \]

where \( s \) represents the non-duplicated portion of the sample covariance matrix (top or bottom half of the matrix) and \( \sigma(\theta) \) is the non-duplicated portion of the model covariance matrix calculated from the model parameters. The residual, \( [s - \sigma(\theta)] \), is the difference between the non-duplicated portions of the sample and model covariance matrices. \( W^{-1} \) is an inverted weight matrix. This is an inverted asymptotic covariance matrix of the sample variances and covariances, as well as the typical elements based on a combination of the second- and fourth-
order product moments of the observed variables (e.g., the sample). The asymptotic covariance weight matrix is estimated using the sample covariance matrix, including the fourth-order (kurtosis) moment, adjusting for violations in multivariate normality in the data.

The WLS fit function is as follows when the metric of the ordinal variables is taken into account:

$$F_{WLS} = ([\hat{\rho} - \sigma(\theta)]' W^{-1} [\hat{\rho} - \sigma(\theta)])$$

where $\hat{\rho}$ is a vector containing the sample tetrachoric or polychoric correlations for the non-duplicated correlations between all pairs of the underlying latent response variables ($y^*$), and $\sigma(\theta)$ is the corresponding vector for the model correlation matrix. $W^{-1}$ is the inverted asymptotic covariance matrix of $\hat{\rho}$. Both the asymptotic covariance matrix and the tetrachoric or polychoric correlations are required to correctly model the response format of ordinal data.

WLS uses a full weight matrix and the asymptotic covariance matrix can lead to practical problems because the inverse of this matrix ($W^{-1}$) must be calculated. As the number of observed variables increases, the number of elements in the W matrix increases resulting in the dimensions to be large (T. A. Brown, 2006; Finney & DiStefano, 2006; Wirth & Edwards, 2007). For example, the W matrix of 20 items would be comprised of 22,155 distinct elements. Simply put, estimating the inverse of the W matrix ($W^{-1}$) is computational intensive and it requires a significantly large sample sizes to avoid positive definite issues. It is recommended that the sample size must be larger than the number of parameters to estimate the $W^{-1}$ accurately (T. A. Brown, 2006; Wirth & Edwards, 2007).

A solution to this problem, especially for smaller samples, is to use a mean- and variance-adjusted WLS estimator, called WLSMV or mean-adjusted WLSM. Unlike WLS, which uses a full-weight matrix, the W matrix for these estimators is diagonal and the number of elements is
equal to the number of correlations between the observed variables in the sample data. Since the inverted W matrix ($W^{-1}$) is the same as the W matrix, the concern of a positive definite W matrix is avoided. WLS, WLSM and WLSMV will produce the same estimate when the sample size is large. To estimate the model parameters, the full-weight matrix for WLS, WLSM and WLSMV is used. In the case of WLSM and WLSMV, the full-weight matrix is a diagonal matrix only, unlike WLS, which is an entire full-weight matrix. In estimating the chi-square model fit and standard error, WLS requires the weight matrix to be inverted, but WLSM and WLSMV does not invert the diagonal weight matrix (Finney & DiStefano, 2006; Wirth & Edwards, 2007).

Extensive research has compared the various estimators under different conditions of normality and variables with different metrics. When the data are continuous and multivariate normal, any NT estimator, including ML (or REML) and GLS, will be accurate. When using ordinal data, the multivariate distribution may occasionally be symmetric and normal looking, but multivariate normality is violated. When the data are categorical or the assumption of multivariate normality is not met, WLS should be used, but only for large samples. If the sample size is small, researchers can utilize WLSMV to prevent the W matrix from being positive definite.

### 2.5.4 Model Specification

In SEM, the structural parameters, variances, error variances, and covariances among the latent and observed variables may either be specified or estimated. Determining which aspects of the model are to be estimated or fixed is called model specification. Bock (1960) showed that CSM and ANOVA models are related in that both can be used to calculate variance components. Bock and Bargmann (1966) introduced CSM as a technique for estimating the variance components of the latent variables for continuous data. Moreover, the authors outlined the
necessary specifications for CSM to be equivalent to a restricted random effects ANOVA model where individuals are the mode of classification.

As outlined by Bock (1960), Bock and Bargmann (1966), and Jöreskog (1978), CSM is specified to meet the assumptions of a restricted random effects ANOVA model for the estimation of variance components. The ANOVA assumes that the factors are uncorrelated and that the variance is equal across each level of the factors. To meet this assumption, the latent variables (known as factors in ANOVA) and error terms are specified to be uncorrelated in the CSM. The error terms are heterogeneous. More specifically, the key concept in using CSM to estimate variance components involves an atypical SEM parameterization. Typically, SEM are formulated to estimate the structural parameters representing the relationship between the observed and latent variables, either by setting the variances of the latent variables to a fixed value or the structural parameters to one. This is often referred to as setting the scale of the latent variable (Brown, 2006). CSM also focuses on the variance of the latent variable(s). As a result, the structural parameters relating the observed to the latent variables are set to the same fixed value (e.g., one), allowing for the variance of the latent variables to be estimated. The option to set the structural parameters and estimate the variance of the latent variables mimics hierarchical linear, random or mixed-effects models that analyze growth and change (see Zumbo, Wu, & Liu, 2013).

Having recognized the importance of model specification and the two covariance matrix methods (Q- and R-), the reader can now appreciate why only a relative G-coefficient is of interest in this dissertation. Given that CSM decomposes the common variance among items (R-method) and is specified to be equivalent to a random effects ANOVA with person as the mode of classification, only the variance due to person (the object of measurement) and all interactions
with it can be estimated. This results in only a relative G-coefficient to be computed, where all factors are considered random. The variance components due to the other factors and their interactions can be estimated by analyzing the matrix of covariances between individuals (Q-method). This mode of analysis is directly analogous to the treatment of data by Bock (1960), Burt (1947), Bock and Bargmann (1966), Jöreskog (1978) and L. Muthén (1983). These authors believed that items should only be correlated based on the dimensions that differentiate among participants (i.e., their general ability). For this reason alone, these authors included only the person facet (object of measurement) and all interactions with it as latent variables to explain the covariance in items.

2.6 Literature Review of CSM with Continuous and Ordinal Data for GT

Bock (1960), Burt (1947), and Bock and Bargmann (1966) discussed continuous variables, or ordinal variables treated as if they were continuous. Following in this tradition, Marcoulides (1996), Raykov and Marcoulides (2006), and Schoonen (2005) demonstrated how to estimate the relative G-coefficients of these variables using CSM for continuous data or ordinal data treated as if it were continuous. These studies confirm that a relative G-coefficient can be computed for crossed (Marcoulides, 1996; Raykov & Marcoulides, 2006) and nested (Schoonen, 2005) designs. L. Muthén (1983) was the first to show that CSM of a tetrachoric correlation matrix is possible in estimating variance components for GT. She compared the variance components, standard errors, model fit indices, reliability and G-coefficients of binary response data from a random effects ANOVA model—where the data are treated as continuous—to CSM—where the data are treated as categorical. She illustrated this using real and simulated data for a one- and two-facet crossed design.
L. Muthén (1983) showed that CSM produced more accurate estimates of the variance components and G-coefficients than did the ANOVA model, and that the reliability of the latter decreased as the number of items or levels in the factors increased. Based on her results, it is reasonable to conjecture that modeling the tetrachoric correlations of the underlying latent response variable leads to unbiased estimates of the variance components and relative G-coefficients than does treating ordinal data as continuous.

The findings from L. Muthén’s (1983) work have vital implications for this dissertation. Firstly, she demonstrates that it is possible to model the tetrachoric correlation matrix of the underlying latent response variables of ordinal data. Secondly, her results suggest that CSM of ordinal data may provide more accurate estimates of G-coefficients than the MS method, which treats the variables as continuous. By extension, this implies that ordinal GT can be done using CSM of the polychoric correlation matrix of polytomous items and nested designs.

Since L. Muthén (1983) wrote her unpublished dissertation, no progress has been made in estimating the reliability of ordinal data in GT, despite the popularity of Likert scales or response scales in most disciplines. Because it was beyond the scope of her work, L. Muthén (1983) did not systematically examine the difference in relative G-coefficients between treating ordinal data as continuous and ordinal when varying the number of response categories, skewness or magnitude of the population reliability values. Consequently, researchers still do not appreciate the benefits of considering the metric of ordinal data for GT.

2.7 Purpose of the Study

Given the dearth of research on modeling the response format of observed variables in GT, the purpose of this dissertation is twofold. Firstly, it seeks to demonstrate that variance components can be estimated for binary and polytomous data in order to calculate a relative G-
coefficient for a crossed and nested design using CSM. To compute the generalizability coefficients, it will rely on the framework set out by L. Muthén (1983) and Zumbo et al. (2007) for modeling the underlying latent response variable of the ordinal data. Secondly, the dissertation will systematically compare the relative G-coefficients and bias of ordinal data treated as continuous (GT) and as ordinal (ordinal GT) using CSM when varying the number of response categories, skewness of the item response distribution, and magnitude of the population’s G-coefficient.

Chapter 3 consists of two studies that explore the aforementioned objectives. In the first study, ordinal GT is demonstrated using CSM of ordinal data for both a crossed and nested design using real data sets. In the second, a Monte Carlo simulation study, a sample of 100,000 participants was generated for each theoretical G-coefficient. This simulation study closely follows the methodology of Zumbo et al. (2007). It manipulates three main conditions: the number of response categories (two to seven), the magnitude of the theoretical G-coefficients (0.40, 0.60, 0.80, and 0.90), and the skewness of the item response distributions (0 and -2). It examines both a two-facet fully crossed and a one-facet partially nested design. The relative G-coefficient and ordinal relative G-coefficient are then compared to the theoretical G-coefficient. The results of this study illustrate the implications of treating ordinal data as continuous in reliability analyses, and offer researchers a guide on the extent of the bias in the coefficients reported in the extant literature.
Chapter 3: Conducting Ordinal Generalizability Theory for Ordinal Response Data

3.1 Introduction

Generalizability theory (GT) provides a framework for measuring the accuracy with which one can generalize from a set of observed scores to a universe of observations (Shavelson & Webb, 1991). It extends classical test theory (CTT) in that it estimates both systematic and random sources of error in one analysis. In GT, observed scores obtained from a measure are considered a sample from a universe of admissible observations, which consists of all possible observations that could be considered interchangeable to those actually obtained. The universe is defined by all possible combinations of the various levels of the facets, or measurement conditions. These measurement conditions are referred to as facets in GT, but are equivalent to factors in linear models and latent variables in Structural Equation Modeling (SEM). The universe of generalization is the levels of the facets in the measurement design over which one wishes to generalize. A universe score is the expected value of the observed score over all observations in the universe of generalization (Shavelson & Webb, 1991).

GT quantifies the various sources of error in a given measurement design by estimating generalizability coefficients (G-coefficients). A G-coefficient estimates the proportion of the total variance that is attributable to the universe score (Hoyt & Melby, 1999). In general, the G-coefficient can be estimated from the variance components as follows:

\[
\frac{\sigma^2_r}{\sigma^2_r + \sigma^2_{\text{residual}}},
\]

where \(\sigma^2_r\) is the variance due to the object of measurement (universe-score variance) and \(\sigma^2_{\text{residual}}\) is the variance due to all other sources of error. The most commonly estimated G-coefficients are
the generalizability coefficient (relative decision) and the index of dependability (absolute decision), each of which reflects the type of decision that will be made from a set of observed scores (Brennan, 2001).

Most measurement specialists writing about GT, including Brennan (2001), Shavelson and Webb (1991), and Cardinet, Johnson and Pini (2010), calculate variance components with the assumption that the item response data are continuous. This is surprising, considering that nearly all of GT is conducted on ordinal data (two- to seven-point response categories). To date, researchers have failed to address this disconnect between accounting for the metric of ordinal variables and the estimation techniques used to compute variance components in GT.

Many statistical methods exist to compute variance components for GT, with only a limited number of techniques that can account for the metric of ordinal data. The most common way to calculate variance components, the mean squares (MS) method, uses a factorial or repeated measures analysis of variance (ANOVA). However, one of the assumptions of an ANOVA is that the dependent variable is continuous. A somewhat forgotten technique that can model the metric of ordinal variables is covariance structural modeling (CSM), which relies on a framework of latent variables to estimate variance components. Bock (1960) and Bock and Bargmann (1966) demonstrated the usefulness of CSM in obtaining the variance components of latent variables for continuous data by highlighting its parallels with a random effects ANOVA.

Other authors have utilized an SEM framework, like CSM, to estimate the reliability of a set of observed scores. Zumbo et al. (2007) demonstrated that factor analysis could be used to compute an ordinal version of coefficient alpha. Marcoulides (1996), Raykov and Marcoulides (2006), and Schoonen (2005) successfully employed CSM to estimate variance components for GT using continuous and ordinal data, although they treated the ordinal data as continuous.
These three studies applied the work of Bock (1960) and Bock and Bargmann (1966) to demonstrate that CSM can be used to estimate variance components in computing the relative G-coefficients for crossed (Marcoulides, 1996; Raykov & Marcoulides, 2006) and nested (Schoonen, 2005) designs. L. Muthén (1983) was the first to correctly model binary ordered response data in the computation of variance components for GT, using CSM to compare the correlations between the underlying latent variables and the observed variables. From her results and those of Zumbo et al. (2007), we can conjecture that treating ordinal data as continuous can lead to substantially deflated relative G-coefficients.

Since the early work of L. Muthén (1983), no research has been conducted on computing the variance components for the GT of polytomous data or more complex assessment designs such as those involving nested facets. Before turning to the analysis of ordinal data, let us briefly review the use of CSM in GT when data are treated as continuous.

3.1.1 Covariance Structural Modeling for GT Using Continuous Data

CSM is a type of SEM that analyzes the covariance among multinormal observed variables to estimate the variance due to the latent variables; however, the means of the observed or latent variables need not be included in the model (Bock & Bargmann, 1966; Bock, 1960; Bollen, 1989). CSM decomposes the variance and covariance of data into common and unique variation. Common variance is the variance of the observed variables explained by one or more of the latent variables, while unique variance can be further divided into error variance (e.g., completely random error) and specific variance, which is attributable to the observed variable, not the latent variable.

Two patterns among the observed variables can be examined in CSM: the Q- and R-methods (McKeown & Thomas, 1988). A Q-method analyzes the covariance between
individuals or groups (e.g., person 1 correlated with person 2) across a sample of observed variables. Researchers use this method to understand the similarities among study participants. An R-method analyzes the covariance between the variables (e.g., item 1 correlated with item 2) across a sample of individuals, and is used to identify patterns of variation in individuals’ characteristics. The R-method is the most common type of covariance pattern analyzed in psychological, educational and behavioural research (McKeown & Thomas, 1988). The type of covariance structure being modeled is important because it dictates which variance components can be estimated for a given assessment design when using CSM for GT.

Burt (1947) and Bock (1960) demonstrated that CSM and ANOVA models are related and that CSM can be used to calculate variance components. Bock and Bargmann (1966) introduced CSM as a parameterization technique for estimating the variance components of the latent variables for continuous observed variables—though they also treated ordinal data as continuous in one of their examples. The authors outlined the necessary model specifications for making CSM equivalent to a restricted random effects ANOVA model, where individuals are the primary mode of classification. They accomplished this by specifying the CSM, as outlined by Bock (1960), Bock and Bargmann (1966) and Jöreskog (1978), in a way that met the assumptions of the restricted random effects ANOVA model. The structural parameters (e.g., factor loadings) are all set to the same fixed value (e.g., one), allowing for the variance of the latent variables to be estimated. Setting the structural parameters to a fixed value and enabling the calculation of the variance of the latent variables is essentially what makes CSM different from typical SEM. Furthermore, the ANOVA assumes that the factors are uncorrelated and that the variance is equal across each level of the factors. Therefore, when using CSM to estimate variance components, the model is specified to have uncorrelated latent variables and error
terms, thereby meeting the assumptions of compound symmetry and allowing the error terms to be heterogeneous.

Given that CSM decomposes the common variance among items (R-method) and is specified to be equivalent to a random effects ANOVA with person as the mode of classification, only the variance due to the object of measurement (participants) and all its interaction terms may be estimated for GT. This results in only a relative G-coefficient to be computed, where all facets are random. This restriction is in part due to the model’s conceptualization by Bock (1960), Burt (1947), Bock and Bargmann (1966), Jöreskog (1978), and L. Muthén (1983), who presupposed that the items should be correlated based solely on the dimensions that separate participants on a given construct. Hence, the only factors that should be estimated are the object of measurement and all its interactions. These factors are included as latent variables in CSM. Although the previously mentioned authors believed these latent variables should account for all covariation among items, other scholars (e.g., Bock & Bargmann, 1966; Jöreskog et al., 1981; Marcoulides, 1996; L. Muthén, 1983) have noted that variance components due to the other facets (e.g., items and occasions) and their interactions can be estimated by analyzing the matrix of covariance among individuals (Q-method).

The following example will be used throughout this section to illustrate the CSM specifications for continuous data. Consider a study where ten individuals (i = 1, 2, …, 10) are randomly selected and scored on four randomly chosen items (j = 1, 2, 3, 4) on two separate occasions (k = 1, 2) using a ten-point response scale that rates their interviewing skills. This is a two-facet crossed design where items (t) and occasions (o) have been randomly selected, resulting in a person (p) x items (t) x occasions (o) design (pto). The object of measurement is p, while the facets of generalization are items and occasions.
GT and CSM have particular notational conventions that are not easily combined when utilizing CSM to describe variance components for GT. In line with L. Muthén (1983) and Marcoulides (1996), GT notation will be used within an SEM notational language (see Table 3.1). For the above example, CSM conventional notation for any observed score is written as \( Y_{ijk} \), with \( i, j \) and \( k \) denoting the \( i^{th} \) person, \( j^{th} \) item and \( k^{th} \) occasion, respectively. In GT, one is interested in the identified sources of variance due to each factor in the model, such as person, item and occasion. Combining the notation from both approaches, an observed score is denoted as \( p_{ito} \), where \( pto \) represents the identified sources of variance and the subscripts represent a particular person, item and occasion. This combination facilitates the integration of the GT notational language into the SEM framework.

**Table 3.1. Combination of GT and SEM conventional notation for Ordinal GT**

<table>
<thead>
<tr>
<th></th>
<th>SEM</th>
<th>GT</th>
<th>Combined Notation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>( i )</td>
<td>( p )</td>
<td>( p_{i} )</td>
</tr>
<tr>
<td>Item</td>
<td>( j )</td>
<td>( t )</td>
<td>( t_{j} )</td>
</tr>
<tr>
<td>Occasions</td>
<td>( k )</td>
<td>( o )</td>
<td>( o_{k} )</td>
</tr>
<tr>
<td>Observed Score</td>
<td>( Y_{ijk} )</td>
<td>( Y_{ijk} )</td>
<td>( Y_{ijk} )</td>
</tr>
</tbody>
</table>

In this design, the components of an observed score (\( Y_{ijk} \)) for the \( i^{th} \) person on the \( j^{th} \) item across the \( k^{th} \) occasion can be expressed as a general linear equation as follows:

\[
Y_{ijk} = \mu + p_{i} + t_{j} + o_{k} + p_{ito} + \epsilon_{ijk},
\]

where \( Y_{ijk} \) = the continuous dependent variable;

\( \mu \) = the grand mean;

\( p_{i} \) = the participant effect;
\( t_j \) = the item effect;

\( o_k \) = the occasion effect;

\( pt_{ij} \) = the interaction between person and item;

\( po_{ik} \) = the interaction between person and occasion;

\( to_{jk} \) = the interaction between item and occasion; and

\( e_{ijk} \) = the error term, which also contains the highest-order interaction.

The following restricted linear equation is analyzed using CSM, given the specifications by Burt (1947) and Bock and Bargmann’s (1966):

\[
Y_{ijk} = \mu + p_i + pt_{ij} + po_{ik} + e_{ijk}.
\]

The CSM corresponding to the restricted linear equation relating the observed variables, random facets (or factors or latent variables) and unique variance is:\(^3\):

\[
Y_{jk} = \Lambda \eta_m + \varepsilon_{jk},
\]

where \( \Lambda \) is a \( p \times m \) matrix in which \( p \) is the number of observed variables and \( m \) is the number of random factors, and \( \eta \) is the \( m \times 1 \) vector of the random factors (e.g., person, person by item and person by occasion effects). In the population of participants, the latent variables are assumed to have a multivariate normal distribution with a mean of zero and a covariance matrix, \( \phi. \varepsilon \) is a \( p \times 1 \) vector of the errors for each observed variable, which are random with respect to items and occasions. Furthermore, \( \varepsilon \) is uncorrelated with \( \eta \) and is normally distributed with a covariance matrix (\( \Theta \)).

The model covariance matrix, \( \Sigma(\theta) \), for \( Y_{jk} \) is:

\(^3\) The i subscript has been left out because CSM models the matrix of relationships among items, not persons.
\[ \Sigma(\theta) = \Lambda \phi \Lambda' + \Theta_\varepsilon, \]

where the meaning of \( \Lambda \) does not change, and \( \phi \) is a square and symmetrical covariance matrix of the latent variable (\( \eta \)). \( \Theta_\varepsilon \) is a diagonal matrix of the measurement error (\( \varepsilon \)), where the error terms are assumed to be uncorrelated, but not homogenous.

The \( \Lambda \) matrix is as follows:

\[
\Lambda = \begin{bmatrix}
1 & 1 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 1 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 1 & 1 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 1 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 1 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 1 & 0 & 1 \\
\end{bmatrix}
\]

where column 1 refers to person (\( p \)), columns 2 through 5 to the interaction between person and item (\( pt_1, pt_2, pt_3, pt_4 \)), and columns 6 and 7 to the interaction between person and occasion (\( po_1 \) and \( po_2 \)). \( \Lambda' \) is the inverse of \( \Lambda \). All structural parameters in \( \Lambda \) are set to one, and the variance and covariance components in \( \phi \) and \( \Theta_\varepsilon \) are estimated.

The \( \phi \) is a diagonal covariance matrix among the random facets. The covariances are set to zero because the latent variables are uncorrelated; this is because the ANOVA assumes zero covariances among the random facets. The diagonal elements of the matrix are equal to each variance component needed to compute a relative G-coefficient (i.e., \( \sigma_p^2, \sigma_{pt1}^2, \sigma_{pt2}^2, \sigma_{pt3}^2, \sigma_{pt4}^2, \sigma_{po1}^2, \sigma_{po2}^2 \)). The matrix is as follows:
\[ \Theta, \text{ a diagonal covariance matrix, contains the variances and covariances among the errors of the latent variables. The covariances are set to zero because the error terms are uncorrelated. This meets the ANOVA’s assumption that the error variances are equal and the covariances among them are zero. The diagonal elements are the variance of } pt_f, e_{jk}, \text{ and are denoted simply as } e_{jk}. \]

For each of the four items across the two occasions,

\[ \Theta_e = \begin{bmatrix}
\sigma_{e12}^2 \\
\sigma_{e21}^2 \\
\sigma_{e31}^2 \\
\sigma_{e41}^2 \\
\sigma_{e12}^2 \\
\sigma_{e22}^2 \\
\sigma_{e32}^2 \\
\sigma_{e42}^2 
\end{bmatrix} \]

Alternatively, a stronger assumption can be made where the diagonal variances are constrained to be equal and the error in the items to be the same (e.g., the unique variance is equal for all items, as opposed to being heterogeneous and thus estimated separately). The path diagram for the example is as follows and can be interpreted based on the specifications set out in each matrix:
Figure 3.1. The path diagram for estimating the variance components in a two-facet fully crossed design (pio) using SEM. All the one-way paths (structural parameters) are fixed to one so that the variance of the latent variables can be estimated.

From the path diagram, each of the variance components can be estimated as follows:

\[
\frac{\sigma_{e11}^2 + \sigma_{e21}^2 + \sigma_{e31}^2 + \sigma_{e41}^2 + \sigma_{e21}^2 + \sigma_{e32}^2 + \sigma_{e42}^2}{n_{\text{residual terms}}} = \sigma_e^2
\]

\[
\sigma_{\text{person}}^2 = \sigma_p^2
\]

\[
\frac{\sigma_{pt1}^2 + \sigma_{pt2}^2 + \sigma_{pt3}^2 + \sigma_{pt4}^2}{n_{\text{items}}} = \sigma_{pt}^2
\]
When the variables are continuous, $\Sigma(\theta)$ is a matrix of Pearson correlation coefficients among the observed variables. However, when the variables are ordinal, $\Sigma(\theta)$ is a matrix of tetrachoric (for binary responses) or polychoric (for more than two response categories) correlations between the latent response variables in the model.

It is important to appreciate that an ANOVA is a statistical technique that decomposes the total variance in a set of observed scores. If a factor that accounts for a significant amount of variance is missing from the ANOVA model, the variance of the incorporated factors, including the error term, will not be estimated correctly. In CSM, not all of the factors that account for a significant amount of variance need to be included because this method decomposes the common variance among the observed variables (e.g., items) and attributes any covariation to a small number of latent variables (which are the factors in the ANOVA model).

Bock (1960) and Bock and Bargmann (1966) believed that the observed variables should only be correlated based on the dimensions that distinguish among participants (e.g., construct, ability). Applying this logic to CSM using a GT framework, only the object of measurement and all its interaction terms are included as latent variables to explain for a significant amount of the covariance in the observed variables. The other sources of error (e.g., facets of generalization) are excluded.

The model fit indices not only assess whether the parameters and the model covariance matrix reproduce the sample covariance matrix, but whether the factors (as latent variables) explain for the covariance in the observed variables. The model may not fit the data if a) the parameters are misspecified, b) the metric of the data are not considered, c) a factor is missing, or
d) too much measurement error exists in the items. The model fit indices capture any discrepancies between the model and the sample covariance matrix in the residual matrix. This is not the same as $\Theta_\varepsilon$, which captures for the unique variances of the observed variables.

The residual matrix is conceptualized as follows:

$$\Sigma_{\text{sample}} = \Sigma_{\text{model}} + \Sigma_{\text{residuals}}$$

where $\Sigma_{\text{sample}}$ is the sample covariance matrix, $\Sigma_{\text{model}}$ is the model covariance matrix, and $\Sigma_{\text{residual}}$ is the residual covariance matrix. This equation can be expanded to unpack the model covariance matrix, which consists of a common and unique component, as follows:

$$\Sigma_{\text{sample}} = \Lambda \phi \Lambda' + \Theta_\varepsilon + \Sigma_{\text{residuals}}$$

To estimate variance components for GT, the object of measurement and all its interaction terms are represented in the model covariance matrix by how the model has been specified. All other sources of error (e.g., items, raters, and their interactions) are captured in the residual covariance matrix because they are not included in the model.

In our parameterization of the model, the model fit indices may indicate whether the object of measurement and all its interaction terms account for a significant amount of the covariance among the observed variables, or whether a factor (such as items or raters) is missing (Bock, 1960). That is, if the object of measurement and all its interaction terms account for most of the covariance in the observed variables, the difference between the model and sample covariance matrices will be close to zero. The model fit indices would suggest that the model fits the data. However, if one of the excluded sources of variation in the model, such as items or raters, it accounts for most of the covariance in the observed variables, the model covariance matrix will not reproduce the sample covariance matrix, resulting in values other than zero in the
residual covariance matrix. This would suggest that the model does not accurately represent the data, suggesting that the current latent variables (e.g., the object of measurement and all its interaction terms) do not account for the covariance in the observed variables. Perhaps one of the other sources of variation, such as the raters or items, may be responsible for the covariance in the observed variables. More importantly, this would imply that the measurement tool does not differentiate among individuals with varying ability and will result in a poor G-coefficient.

In summary, the covariance matrix in CSM represents the covariance among the items (R-method). The factors (e.g., latent variables) that are chosen in the model are those that are hypothesized to account for the covariance in the observed variables, which according to Bock (1960), Burt (1947) and Bock and Bargmann (1966), is the latent variables related to participants’ ability. In GT, this would be equivalent to the object of measurement and all other sources of variation that interact with it because the goal of the measure is to differentiate among participants on some construct of interest. Furthermore, the model in CSM is specified to be analogous to a restricted random effects ANOVA with individuals as the mode of classification. Variance components for the other factors (item and occasion) and their interaction (io) can be estimated by modeling the covariance among individuals rather than among items (Bock & Bargmann, 1966; Burt, 1947). In this case, one of the facets of generalization would become the object of measurement in explaining for the covariance among individuals.

Although Bock (1960), Bock and Bargmann (1966) and Jöreskog (1978) outlined the CSM specifications for estimating variance components for continuous data, the same restrictions can be applied for ordinal data. Indeed, L. Muthén (1983) demonstrated that variance components and a relative G-coefficient can be computed using CSM of the tetrachoric correlation matrix. However, she did not a) model polytomous data using the polychoric
correlation matrix, or b) systematically compare the G-coefficients when ordinal data are treated as either continuous or ordinal in CSM by varying the number of response categories, skewness of the item response distributions, and magnitude of the population reliability.

This chapter will examine these limitations. The first study will demonstrate ordinal GT using CSM for both a crossed and nested design of binary and polytomous items in real data sets. In the second study, a Monte Carlo simulation of the population analogue for each theoretical G-coefficient is generated. Three main conditions are manipulated: the number of response categories (two to seven), the magnitude of the theoretical G-coefficients (0.40, 0.60, 0.80, 0.90), and the amount of skewness of the item response distribution (0 and -2). The study investigates both a two-facet fully crossed and a one-facet partially nested design. The purpose of the second study is to examine the difference between treating ordinal data as continuous and ordinal in the estimation of G-coefficients.

3.2 Study 1: Ordinal GT: Covariance Structural Modeling for GT Using Ordinal Data

It is assumed that for each ordinal variable, y, there is an underlying continuous variable called \( y^* \), which is normally distributed with a specified mean and variance (Bollen, 1989; Jöreskog, 1994; B. Muthén, 1984). The points that divide the continuous latent response variable into discrete categories (c) are called thresholds (\( \tau_c \)). The total number of thresholds will equal the number of categories minus one (c-1). These thresholds are not the same as the response or Likert categories. The thresholds of the underlying latent variable can be spaced at equal or unequal intervals and can be thought of as estimating the response-scale cutoffs of the continuous latent variable.

Formally, the connection of y for item j with c number of response categories to the underlying latent variable \( y^* \) can be written as:
\[ y_j = c \text{ if } \tau_c < y_j^* < \tau_{c+1}, \]

where \( \tau_c \) and \( \tau_{c+1} \) are the thresholds on the underlying continuum of the latent variable. The thresholds can be spaced at equal or unequal intervals, satisfying the constraint:

\[-\infty < \tau_0 < \tau_1 < \ldots < \tau_{c-1} < \tau_c < \infty.\]

When the variable responses are ordinal and are treated as such, the covariance matrix is estimated from a series of tetrachoric or polychoric correlations. A tetrachoric correlation is used to represent the relationship between binary data (e.g., yes/no or 0/1), whereas a polychoric reflects the relationship between polytomous data (i.e., three or more response categories). The tetrachoric and polychoric correlations are estimated from the bivariate distributions of each threshold between two latent continuous variables, \( y_1^* \) and \( y_2^* \), of the corresponding ordinal variables, \( y_1 \) and \( y_2 \). These correlations are among the underlying latent variables of the ordinal variables. The use of these thresholds in the estimation of the sample covariance matrix is much more precise than using the actual ordinal response or Likert categories (e.g., 1, 2, 3, 4 or 5).

When using tetrachoric and polychoric correlations, it is assumed that the underlying latent response variable is continuous and linearly related to the dependent variable (Holgado–Tello et al., 2010).

There are a number of theoretical and empirical reasons the Pearson correlation matrix should be avoided to examine the degree of association between ordinal variables. Theoretically, ordinal variables imply ordinal scales and thus require statistical analyses that can account for the ordinal metric. The Pearson correlation, however, assumes continuous data. Moreover, this method typically reduces the variability and covariance of ordinal data (Holgado–Tello et al., 2010), which may produce biased parameter estimates when using SEM techniques. Empirically, a large number of studies show that when the metric of ordinal data are ignored, the parameter
estimates, hypothesis testing, variance components and reliability results are inaccurate (Gadermann et al., 2012; Holgado–Tello et al., 2010; Jaeger, 2008; L. Muthén, 1983; Rhemtulla et al., 2012; Zumbo et al., 2007). At least five response categories seem to be required for the Pearson correlation to produce similar estimates as the tetrachoric or polychoric correlations, though this may be untrue when the data are extremely skewed.

Recently, Zumbo, Gadermann, and Zeisser (2007) developed ordinal coefficient alpha and theta as a new method for computing coefficient alpha when item responses are ordinal. They demonstrated that either factor analysis (ordinal coefficient alpha) or principle components analysis (ordinal coefficient theta) may be used to model the tetrachoric and polychoric correlation matrix of the underlying latent variable of ordinal data. In this and a follow-up study (Gadermann et al., 2012), the researchers found that coefficient alpha consistently produced a negatively biased estimate of the actual reliability for two to seven response categories. This was especially true when the item response distributions were skewed. Ordinal alpha, by contrast, always reflected an unbiased estimate of the true reliability for all response categories, even with skewed item distributions. These findings suggest that the metric of ordinal data should not be ignored when computing coefficient alpha and that G-coefficients may fair no different when the ordinal metric is ignored.

In an unpublished dissertation, L. Muthén (1983) demonstrated that CSM of the tetrachoric correlation matrix is possible in estimating variance components for GT. She compared the variance components, standard errors, model fit indices and relative G-coefficients of binary response data from a random effects ANOVA, where the data are treated as continuous, to CSM, where the data are treated as ordinal. By analyzing a one- and two-facet crossed design using real and simulated data, she found that CSM, which accounted for the
metric of the ordinal data, produced more accurate estimates of the variance components and relative G-coefficients than did the ANOVA model. Moreover, the difference in the variance components and relative G-coefficients between the two analyses decreased as the number of items or values of the levels in the other facets increased. Her results suggest that modeling the tetrachoric correlations of the underlying variable leads to more accurate and unbiased estimates of the variance components and relative G-coefficients.

Marcoulides, L. Muthén, and Zumbo et al. all use SEM to estimate reliability indices. Zumbo et al. (2007) use factor and principle components analysis to calculate coefficient alpha and theta, while Marcoulides (1996) and L. Muthén (1983) rely on CSM to estimate G-coefficients. The major difference in the work of these researchers is that L. Muthén (1983) and Zumbo et al. (2007) model the underlying continuous latent variable of the ordinal data to estimate reliability indices, which allows them to account for the metric of ordinal data. Marcoulides (1996) simply applies the notions of Bock (1960) and Bock and Bargmann (1966) by utilizing CSM to compute G-coefficients.

This study extends L. Muthén’s (1983) work by demonstrating that variance component estimation for a nested design and polytomous data are also possible when modeling the polychoric correlation matrix using CSM. Modeling the underlying latent variable of the ordinal data for GT will be referred to as ordinal GT. This study will also develop a syntax of computing variance components and G-coefficients for ordinal data using LISREL (Jöreskog & Sörbom, 2006) and Mplus (Muthén, B., & Muthén, L., n.d.).
3.2.1 Methods: Demonstration of Covariance Structural Modeling for GT on a Tetrachoric and Polychoric Matrix Using Ordinal Data

*Study Design*

The study uses two example data sets from a GT software package called EduG (Cardinet et al., 2010). The variance components and relative G-coefficients reported by EduG are calculated using the MS method and will be referred to as the G-coefficient (MS). Restricted maximum likelihood (REML) is conducted via the traditional MS method because it provides better estimates of variance components in data with normal and non-normal distributions and in balanced and unbalanced ANOVA designs (Marcoulides, 1990). The goal of this study is to determine whether it is feasible to compute variance components using CSM when the analysis treats polytomous data as ordinal. This will be demonstrated in both a crossed and a nested design.

The study first replicates the variance components and relative G-coefficients reported by EduG for a fully crossed and a partially nested data set when the ordinal data are treated as continuous using CSM. This model will be referred to as GT (CSM) and the estimated G-coefficients as relative G-coefficients. LISREL and Mplus are used to compute the variance components and G-coefficients (CSM). For LISREL, a slightly modified version of Marcoulides’ (1996) syntax is used for the crossed design and a highly modified version of Schoonen’s (2005) syntax for the nested design. The LISREL syntax for both of these designs can be found in Appendix 1. The Mplus syntax for the study designs was created by the author of this paper and can be found in Appendix 2. Once the author has demonstrated that GT (CSM) can replicate the MS method’s results for EduG’s data sets, the same analyses will be conducted using the tetrachoric and polychoric correlation matrix in Mplus for ordinal GT. Ordinal GT will be
referred to as ordinal GT (CSM) and the G-coefficients as the ordinal relative G-coefficients. Ordinal GT was modeled by changing the syntax in LISREL and Mplus to use the tetrachoric and polychoric correlation matrix instead of the Pearson correlation matrix.

To estimate the variance components for a relative G-coefficient (CSM), unweighted least squares (ULS), and maximum likelihood (ML) are used in LISREL, while ML is used in Mplus\(^4\). Weighted least squares (WLS) or mean- and variance-adjusted weighted least squares (WLSMV) in Mplus are used to obtain the variance components for the ordinal G-coefficients (CSM). The sample size and the number of parameters to be estimated determined whether to use WLS or WLSMV (Brown, 2006; Finney & DiStefano, 2006; Wirth & Edwards, 2007). The WLS method requires the minimum sample size to be greater than the number of estimated parameters (including thresholds). In general, WLS should not be used with small samples and a large number of parameters to be estimated because the risk of a non-positive definite weight matrix is high; WLSMV is therefore used instead.

This study seeks to prove that using ULS and Marcoulides’ syntax in LISREL can produce the same variance components and relative G-coefficient (CSM) as EduG using the MS method for each respective data set. It then attempts to show that ML in LISREL and Mplus can generate the same relative G-coefficients (CSM), providing future justification for using Mplus in computing ordinal relative G-coefficients (CSM). The decision in using Mplus over LISREL is because it implements more general algorithms and can simultaneously estimate the thresholds and polychoric correlations between the items, allowing CSM to be conducted in one simple step. Moreover, it is important to note that the goal of this study is not to compare the results from the different software programs, but rather, to contrast the values obtained from treating

\(^4\) Due to the Mplus default programming, ULS cannot be used when all variables are treated as continuous. As a result, this method cannot be employed.
ordinal data as continuous and ordinal in CSM within each software package with those reported by EduG. The estimated variance components and relative G-coefficients are not expected to be the same across the different statistical programs even when using the same or an equivalent estimator, because each employs different algorithms to implement the estimators. Any discrepancies among the results are more likely to be due to the algorithm in each program than meaningful differences.

Description of the Data Sets

Two example data sets were taken from EduG (Cardinet et al., 2010). The two-facet fully crossed design consists of 120 participants (p), who were scored by two raters (r) on four separate ability domains (d) on a four-point response scale for a total of 960 data points. The notation for the design is prd. The sources of variance are p, r, d, pr, pd, rd, and e, where the error term is confounded by the higher-order interaction term (prd). All but one of the items was negatively skewed (the skewness ranged from 0.15 to -1.89, with one of the items being -2.85).

The one-facet nested design consists of 100 participants (p) who received scores on four math items (i) on three different versions (v) of a test. Four different items were used in each version of the math test, resulting in a total of 12 items, which were graded as zero (incorrect) or one (correct). This leads to a total of 1200 data points. The notation of this design is p(i:v). The sources of variance are p, pv, i:v, and e, where the error term is confounded by the interaction term (p(i:v)). None of the item distributions was skewed (skewness ranged from -1.19 to 1.19). In both example data sets, the object of measurement is the participant and the facets of generalization are items, version, raters and domains.
Data Analysis

Participants, items and versions have been chosen randomly, and are therefore considered random effects in the models. Although domains and raters could be treated as fixed effects, they will be treated as random because only a relative G-coefficient is of interest. The relative G-coefficients were computed using the following equation for the crossed designs:

\[
E \rho^2 = \frac{\hat{\sigma}_p^2}{\sigma_p^2 + \frac{\hat{\sigma}_{pv}^2}{n_v} + \frac{\hat{\sigma}_{pd}^2}{n_d} + \frac{\hat{\sigma}_{prd,e}^2}{n_r n_d}}
\]

and the following equation for the nested ones:

\[
E \rho^2 = \frac{\hat{\sigma}_p^2}{\sigma_p^2 + \frac{\hat{\sigma}_{pv}^2}{n_v} + \frac{\hat{\sigma}_{pi,v,e}^2}{n_i n_v}}
\]

The variance components and relative G-coefficients were compared when ordinal data were treated as continuous and ordinal using CSM within each software program and to the findings from EduG. The results presented are only for a relative G-coefficient.

3.2.2 Proof of Concept: Results

The variance components and G-coefficients for GT (CSM) and ordinal GT (CSM) of the different estimators, along with the reported MS method values from EduG, are presented in

---

5 For the sake of brevity, the term “relative” will no longer be used. The reader should assume that a G-coefficient is for a relative decision unless informed otherwise.
Table 3.2 for the two-facet fully crossed design and Table 3.3 for the one-facet partially nested design. Because REML produces identical results to the MS method for both designs, it is not included as a separate column.

**Table 3.2 Variance Components and relative G-coefficients for EduG's (n = 120) two-facet fully crossed design (prd)**

<table>
<thead>
<tr>
<th>Sources of Variance</th>
<th>Variables treated as continuous</th>
<th>Variables treated as ordinal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EduG (ULS)</td>
<td>LISREL (ML)</td>
</tr>
<tr>
<td>Person</td>
<td>0.02</td>
<td>0.03</td>
</tr>
<tr>
<td>Person x Rater</td>
<td>0.13</td>
<td>0.13</td>
</tr>
<tr>
<td>Person x Domains</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>Error</td>
<td>0.17</td>
<td>0.17</td>
</tr>
<tr>
<td>Relative G-coefficient</td>
<td>0.19</td>
<td>0.19</td>
</tr>
</tbody>
</table>

**Table 3.3 Variance components and relative G-coefficients for EduG’s (n = 100) one-facet nested design (p(i:v))**

<table>
<thead>
<tr>
<th>Sources of Variance</th>
<th>Variables treated as continuous</th>
<th>Variables treated as ordinal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EduG (ULS)</td>
<td>LISREL (ML)</td>
</tr>
<tr>
<td>Person</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>Person x Version</td>
<td>0.02</td>
<td>0.03</td>
</tr>
<tr>
<td>Error</td>
<td>0.20</td>
<td>0.20</td>
</tr>
<tr>
<td>Relative G-coefficient</td>
<td>0.66</td>
<td>0.64</td>
</tr>
</tbody>
</table>

When ordinal data are treated as continuous for the fully crossed and partially nested designs, the results from LISREL are identical to those reported by EduG when using ULS.

However, ML in both LISREL and Mplus did not reproduce the results reported in EduG. This is not surprising considering ML is not equivocal to ULS or the MS method. LISREL and Mplus

---

6 While three decimal places were used in the calculation of the G-coefficients, only two were reported as per APA guidelines.
produced identical G-coefficients (CSM). In general, the magnitude of the ordinal G-coefficient (CSM) is greater than both the G-coefficients (CSM) and the MS method (including REML), in the crossed designs and in the nested design.

3.2.3 Discussion

This study demonstrates that CSM using ULS produces exactly the same variance components and G-coefficients as the MS method when the ordinal data are treated as continuous. This is similar to what Marcoulides (1996), Raykov and Marcoulides (2006), and Schoonen (2005) found when they treated ordinal data continuously. The difference between ML in LISREL and Mplus in reproducing the MS method results reported by EduG is not surprising. ML is not the MS method, as the likelihood function is being optimized rather than the residual sums of the squares being minimized. Furthermore, the data sets had small sample sizes and the item distributions for one of the items was skewed. Under these conditions, ML produces less stable parameter estimates (e.g., factor loadings) than those produced by other estimators such as WLS.

More importantly, this study demonstrates that estimating variance components and ordinal GT is possible using CSM of the polychoric correlation matrix. By modeling the polychoric or tetrachoric correlation matrix, the estimated G-coefficient was greater than when treating the ordinal data as continuous. This suggests that modeling ordinal data using a Pearson correlation matrix may underestimate the relationship among the variables, potentially resulting in an inaccurately low G-coefficient. These results replicate those of scholars who utilize other SEM techniques, such as factor analysis, where using Pearson correlations for ordinal data can significantly underestimate the relationship between ordinal variables (Holgado–Tello et al., 2010). The negatively biased G-coefficients obtained when ordinal data are treated as continuous
are similar to those obtained by Zumbo et al. (2007) and Gadermann et al. (2012) for ordinal alpha, as well as by L. Muthén (1983) for relative G-coefficients. Both the nested and crossed designs demonstrated ordinal G-coefficients were greater in magnitude than G-coefficients. To prove that ordinal GT leads to more accurate G-coefficient values, a simulation (Study 2) will be presented. It will show that ordinal GT is feasible in both nested and crossed designs and that the ordinal G-coefficients are greater in magnitude than regular G-coefficients.

One limitation of the model parameterization is that only the variance due to the object of measurement and the factors that interact with it can be estimated. This, in turn, means one can compute only relative G-coefficients. While researchers can partly resolve the problem by using a Q-method, this can be cumbersome because they must perform a separate analysis for each facet of generalization, with the object of measurement becoming columns in the data set. If people are the object of measurement—which most often is the case in psychological and educational research—then the number of columns is contingent on the sample size. For example, if there are 100 participants, each of which provides a response on four items on two separate occasions, then the number of columns to determine the variance due to occasion with all interactions associated with it would be 400. As one can see, switching from an R- to a Q-method can quickly become laborious. Unless a researcher wishes to conduct a separate CSM using a Q-method for each facet of generalization, he/she may be limited to computing a relative G-coefficient.

Although the popularity of GT is increasing, very little research has been conducted on advancing the statistical techniques required to compute the variance components of ordinal data. This study confirms that ordinal GT can be conducted through CSM using the tetrachoric and polychoric correlation matrix in both crossed and nested designs. Although L. Muthén (1983)
proved this was true for binary response data, the present work is the first to show that ordinal GT is possible for polytomous item responses with complex (e.g., nested) designs. However, while it has been demonstrated that variance components can be estimated and a relative G-coefficient computed while considering the metric of ordinal data using CSM, the benefits of this method are still not appreciated compared to treating ordinal data as continuous.

Study 2 will systematically examine ordinal GT and its statistical impact on G-coefficients using the methodology of Zumbo et al. (2007). The simulation will determine which G-coefficient—ordinal or continuous—more accurately estimates the true reliability in a population. It will allow us to establish the bias in the G-coefficient for different response categories when the analysis treats ordinal data as either continuous or ordinal. This will help determine for which response categories, if any, ordinal GT and traditional GT produce the same G-coefficient value. This will offer empirical results on when and why ordinal data should be treated as ordinal when computing G-coefficients.

3.3 Study 2: A Monte Carlo Simulation Documenting the Effects of Ordinal G-coefficients

3.3.1 Introduction

The previous study demonstrated that ordinal GT is possible using the tetrachoric and polychoric correlation matrix for publically available data. A Monte Carlo simulation will be conducted to compare the G-coefficients of ordinal data treated as continuous and ordinal using CSM, varying the number of response categories, skewness of the item response distributions and magnitude of the population’s (theoretical) G-coefficient. The study calculates the G-coefficients for a two-facet fully crossed and one-facet partially nested design. It also finds the population bias—computed as a percent attenuation—of both the ordinal and traditional G-
coefficients. The experimental manipulations and methodology of this study closely follow those presented in Zumbo et al. (2007).

3.3.2 Methods

Study Design

The data were generated using a Monte Carlo simulation in Mplus to reflect four theoretical G-coefficients: 0.40, 0.60, 0.80 and 0.90, in a population of \( n = 100,000 \). The large sample size is intended to provide population analogues of the coefficients. This study manipulates three experimental factors:

1. The magnitude of the theoretical G-coefficients (0.40, 0.60, 0.80 and 0.90)
2. The number of response categories of the ordinal data (2, 3, 4, 5, 6 and 7)
3. The amount of skewness (normality\(^7\) of the distribution, 0, -2)

The manipulation of these factors will result in a total of 48 conditions. A typical crossed and nested design is simulated to represent the most common research designs encountered in practice. In the crossed design, participants (p) are scored on four items (i) by two different raters (r), resulting in a pri design. In the nested design, participants (p) are rated at three different times (t) on two items (i) by two different raters (r) at each point. This results in a total of six raters, with two raters nested in each time point. The notation for this design is \( p_i(r:t) \). The goal of the simulation study is to compare the percent attenuation of ordinal G-coefficients and G-coefficients under different response categories and skewness for both crossed and nested

\(^7\) It should be noted that the standard practice in the literature of describing ‘bell-shaped symmetric ordinal data’ as ‘normal’ is followed even though ordinal data, by definition, is not normal.
designs. This study uses WLS to analyze the data sets in estimating the variance components for the computation of G-coefficients\(^8\).

**Data Generation**

In each of the 48 conditions, separate data sets were generated using the following factor analytic model written in matrix notation for the crossed design:

\[
Y_{jk} = \Lambda_{1y} \eta_p + \Lambda_{2y} \eta_{pt} + \Lambda_{3y} \eta_{pr} + e_i.
\]

For the nested design, the model is written as follows:

\[
Y_{ji} = \Lambda_{1y} \eta_p + \Lambda_{2y} \eta_{pi} + \Lambda_{3y} \eta_{pr} + e_i.
\]

The crossed factor analytic model was used to generate eight independent ordinal item response vectors, while the nested factor analytic model was used to generate 12 for a sample size of 100,000 participants. The values of the response scales points varied between two and seven depending on the condition. This resulted in a multivariate distribution along eight dimensions for the crossed design and 12 dimensions for the nested design. Ordinal data were generated by setting the thresholds of the underlying item response distributions of the factor analytic model to predetermined thresholds (see Appendix 3). The factor loadings were set to one and the variance due to each of the facets (\(\eta_{item}\), \(\eta_{time}\) and \(\eta_{rater}\)) in the model, including the object of measurement (\(\eta_{person}\)) and the error term (\(e\)) for each item, were set to fixed values to obtain G-coefficients of 0.40, 0.60, 0.80 and 0.90 (see Appendix 4 for the system of latent variable regression equations for each design). The obtained parameters were similar, to 3 decimal points, to those specified in the model, indicating that the algorithm generated the intended data accurately and had a high parameter recovery. The simulation algorithm was examined using two parameter recovery

---

\(^8\) Due to the large sample size, WLS produced the same results as WLSM or WLSMV (Finney & DiStefano, 2006; Wirth & Edwards, 2007).
methods (comparing the model-specified parameters to the generated data, and log-likelihood). The parameter bias between the model-specified and actual values was less than 5% and the log-likelihood was also less than 5%.

Outcome Variables and Data Analysis

For each of the 48 conditions, CSM is conducted on the item response to estimate the variance due to the object of measurement (person), all interaction terms associated with it, and the error term, using each simulated data set. With the resulting variance components, the G-coefficient and ordinal G-coefficients are computed for a relative decision in each of the 48 conditions for the crossed and nested designs. The sources of variance for the crossed design are p, pi, pr and e, and for the nested design, p, pi, pt and e. The G-coefficients are calculated using the following equation for the crossed designs:

$$E \rho^2 = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{pi}^2}{n_i} + \frac{\sigma_{pr}^2}{n_r} + \frac{\sigma_{pir,e}^2}{n_in_r}}$$

and for the nested:

$$E \rho^2 = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{pi}^2}{n_i} + \frac{\sigma_{pt}^2}{n_t} + \frac{\sigma_{pir,e}^2}{n_in_tn_r}}$$

The degree to which ordinal GT and GT accurately estimate the theoretical G-coefficient is determined as the percentage of attenuation for each of the 48 conditions as follows:

$$\frac{(\text{G-coefficient} - \text{theoretical G-coefficient})}{\text{theoretical G-coefficient}} \times 100$$
The G-coefficient denotes either an ordinal G-coefficient or a G-coefficient. The percentage of attenuation or bias ranges from +100 to -100, assuming that the theoretical value is greater than zero. When either the ordinal G-coefficient or the G-coefficient is equal to the theoretical G-coefficient, then the attenuation is zero, which suggests there is no bias. When the percent attenuation is positive, the ordinal G-coefficient or G-coefficient is an overestimate of the theoretical G-coefficient, while a negative value indicates an underestimate.

3.3.3 Results

The G-coefficients for the simulated data sets are presented in tables 3.5 to 3.8 for the crossed design and tables 3.9 to 3.12 for the nested design across the theoretical G-coefficients of 0.40, 0.60, 0.80 and 0.90 by the number of response categories (2 to 7) and level of skewness (0 and -2). Figures 3.2 to 3.5 illustrate the percentage of bias for the crossed design and figures 3.6 to 3.9 for the nested design. Regardless of the design, when the ordinal data are treated as continuous, the G-coefficient is consistently an underestimate of the theoretical G-coefficient value. This ranges from 18% to 0.3% when item responses are normally distributed. When the responses are negatively skewed, the negative bias ranges from 29% to 5%. This bias becomes more evident with binary response categories and a theoretical G-coefficient of 0.40. The bias is less prevalent as the number of response categories and magnitudes of the theoretical G-coefficient increase in value.

In contrast, the ordinal G-coefficients are mainly unbiased (less than 1% on average). Surprisingly, the ordinal G-coefficients for all response categories have a consistent positive bias for the nested design when the item responses are normally distributed or skewed and for the crossed design only for skewed item responses. This results in an overestimation of the theoretical G-coefficient. Although this may appear alarming, the highest overestimated positive
bias is less than 3% with a binary response scale and a theoretical G-coefficient of 0.40. For all other response categories and theoretical G-coefficients, the positive bias ranges from 1.4% to 0.7%. Only for the crossed design with normally distributed item responses is the ordinal G-coefficient consistently unbiased or slightly negatively biased. This negative bias is most evident with binary responses and a theoretical G-coefficient of 0.40. A trend can be observed: binary responses and theoretical G-coefficients of 0.40 result in the most biased G-coefficient estimates regardless of whether the responses are treated as ordinal or continuous.

The G-coefficient approaches the theoretical value as the response categories increases and the theoretical G-coefficient approaches one. In the crossed design for when the item response distributions are not skewed, the G-coefficient approaches the theoretical value after the five-point response scale. In the nested design, the G-coefficient approaches the theoretical value after the four-point response scale. For all other situations, the G-coefficient increases as the theoretical value increases when the data are not skewed.

Based on the data, continuous CSM dramatically underestimates the variance due to person and all its interactions, while consistently inflating the error variance. However, in ordinal CSM, the estimated variance components are consistently closer to the true values.

### Table 3.4. Relative G-coefficients for the theoretical G-coefficient of 0.40 for the crossed design

<table>
<thead>
<tr>
<th># of scale points</th>
<th>0</th>
<th>Skewness</th>
<th>-2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>G-coefficient</td>
<td>Ordinal G-coefficient</td>
<td>G-coefficient</td>
</tr>
<tr>
<td>2</td>
<td>0.327</td>
<td>0.394</td>
<td>0.285</td>
</tr>
<tr>
<td>3</td>
<td>0.362</td>
<td>0.394</td>
<td>0.309</td>
</tr>
<tr>
<td>4</td>
<td>0.376</td>
<td>0.397</td>
<td>0.306</td>
</tr>
<tr>
<td>5</td>
<td>0.384</td>
<td>0.397</td>
<td>0.319</td>
</tr>
<tr>
<td>6</td>
<td>0.386</td>
<td>0.396</td>
<td>0.327</td>
</tr>
<tr>
<td>7</td>
<td>0.389</td>
<td>0.396</td>
<td>0.336</td>
</tr>
</tbody>
</table>
Table 3.5. Relative G-coefficients for the theoretical G-coefficient of 0.60 for the crossed design

<table>
<thead>
<tr>
<th># of scale points</th>
<th>Skewness</th>
<th>Ordinal G-coefficient</th>
<th>0</th>
<th>G-coefficient</th>
<th>Ordinal G-coefficient</th>
<th>-2</th>
<th>G-coefficient</th>
<th>Ordinal G-coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td></td>
<td></td>
<td>0.514</td>
<td>0.598</td>
<td>0.470</td>
<td>0.603</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>0.556</td>
<td>0.597</td>
<td>0.500</td>
<td>0.602</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td>0.575</td>
<td>0.598</td>
<td>0.496</td>
<td>0.602</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td>0.583</td>
<td>0.598</td>
<td>0.512</td>
<td>0.602</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td>0.588</td>
<td>0.598</td>
<td>0.520</td>
<td>0.601</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td>0.590</td>
<td>0.598</td>
<td>0.530</td>
<td>0.602</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.6. Relative G-coefficients for the theoretical G-coefficient of 0.80 for the crossed design

<table>
<thead>
<tr>
<th># of scale points</th>
<th>Skewness</th>
<th>Ordinal G-coefficient</th>
<th>0</th>
<th>G-coefficient</th>
<th>Ordinal G-coefficient</th>
<th>-2</th>
<th>G-coefficient</th>
<th>Ordinal G-coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td></td>
<td></td>
<td>0.714</td>
<td>0.800</td>
<td>0.674</td>
<td>0.801</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>0.752</td>
<td>0.800</td>
<td>0.702</td>
<td>0.801</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td>0.774</td>
<td>0.800</td>
<td>0.699</td>
<td>0.800</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td>0.781</td>
<td>0.799</td>
<td>0.719</td>
<td>0.800</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td>0.787</td>
<td>0.800</td>
<td>0.728</td>
<td>0.801</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td>0.791</td>
<td>0.799</td>
<td>0.740</td>
<td>0.801</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.7. Relative G-coefficients for the theoretical G-coefficient of 0.90 for the crossed design

<table>
<thead>
<tr>
<th># of scale points</th>
<th>Skewness</th>
<th>Ordinal G-coefficient</th>
<th>0</th>
<th>G-coefficient</th>
<th>Ordinal G-coefficient</th>
<th>-2</th>
<th>G-coefficient</th>
<th>Ordinal G-coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td></td>
<td></td>
<td>0.822</td>
<td>0.900</td>
<td>0.790</td>
<td>0.899</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>0.845</td>
<td>0.900</td>
<td>0.814</td>
<td>0.899</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td>0.871</td>
<td>0.900</td>
<td>0.816</td>
<td>0.899</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td>0.882</td>
<td>0.900</td>
<td>0.836</td>
<td>0.900</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td>0.888</td>
<td>0.900</td>
<td>0.845</td>
<td>0.900</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td>0.891</td>
<td>0.900</td>
<td>0.856</td>
<td>0.900</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 3.8. Relative G-coefficients for the theoretical G-coefficient of 0.40 for the nested design

<table>
<thead>
<tr>
<th># of scale points</th>
<th>Skewness</th>
<th>0</th>
<th>Ordinal G-coefficient</th>
<th>G-coefficient</th>
<th>Ordinal G-coefficient</th>
<th>-2</th>
<th>Ordinal G-coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>G-coefficient</td>
<td></td>
<td>Ordinal G-coefficient</td>
<td>G-coefficient</td>
<td>Ordinal G-coefficient</td>
<td></td>
<td>Ordinal G-coefficient</td>
</tr>
<tr>
<td>2</td>
<td>0.358</td>
<td></td>
<td>0.403</td>
<td>0.322</td>
<td>0.412</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.385</td>
<td></td>
<td>0.398</td>
<td>0.319</td>
<td>0.409</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.398</td>
<td></td>
<td>0.404</td>
<td>0.321</td>
<td>0.410</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.398</td>
<td></td>
<td>0.403</td>
<td>0.327</td>
<td>0.409</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.400</td>
<td></td>
<td>0.402</td>
<td>0.334</td>
<td>0.407</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.400</td>
<td></td>
<td>0.402</td>
<td>0.339</td>
<td>0.408</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 3.9. Relative G-coefficients for the theoretical G-coefficient of 0.60 for the nested design

<table>
<thead>
<tr>
<th># of scale points</th>
<th>Skewness</th>
<th>0</th>
<th>Ordinal G-coefficient</th>
<th>G-coefficient</th>
<th>Ordinal G-coefficient</th>
<th>-2</th>
<th>Ordinal G-coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>G-coefficient</td>
<td></td>
<td>Ordinal G-coefficient</td>
<td>G-coefficient</td>
<td>Ordinal G-coefficient</td>
<td></td>
<td>Ordinal G-coefficient</td>
</tr>
<tr>
<td>2</td>
<td>0.547</td>
<td></td>
<td>0.599</td>
<td>0.511</td>
<td>0.608</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.588</td>
<td></td>
<td>0.602</td>
<td>0.515</td>
<td>0.607</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.594</td>
<td></td>
<td>0.601</td>
<td>0.515</td>
<td>0.603</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.597</td>
<td></td>
<td>0.602</td>
<td>0.524</td>
<td>0.608</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.595</td>
<td></td>
<td>0.602</td>
<td>0.531</td>
<td>0.605</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.598</td>
<td></td>
<td>0.601</td>
<td>0.536</td>
<td>0.605</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 3.10. Relative G-coefficients for the theoretical G-coefficient of 0.80 for the nested design

<table>
<thead>
<tr>
<th># of scale points</th>
<th>Skewness</th>
<th>0</th>
<th>Ordinal G-coefficient</th>
<th>G-coefficient</th>
<th>Ordinal G-coefficient</th>
<th>-2</th>
<th>Ordinal G-coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>G-coefficient</td>
<td></td>
<td>Ordinal G-coefficient</td>
<td>G-coefficient</td>
<td>Ordinal G-coefficient</td>
<td></td>
<td>Ordinal G-coefficient</td>
</tr>
<tr>
<td>2</td>
<td>0.720</td>
<td></td>
<td>0.801</td>
<td>0.669</td>
<td>0.802</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.752</td>
<td></td>
<td>0.801</td>
<td>0.691</td>
<td>0.802</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.783</td>
<td></td>
<td>0.801</td>
<td>0.689</td>
<td>0.802</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.793</td>
<td></td>
<td>0.802</td>
<td>0.711</td>
<td>0.802</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.797</td>
<td></td>
<td>0.801</td>
<td>0.726</td>
<td>0.803</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.799</td>
<td></td>
<td>0.801</td>
<td>0.738</td>
<td>0.803</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3.11. Relative G-coefficients for the theoretical G-coefficient of 0.90 for the nested design

<table>
<thead>
<tr>
<th># of scale points</th>
<th>Skewness</th>
<th>0</th>
<th>0</th>
<th>-2</th>
<th>-2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>G-coefficient</td>
<td>Ordinal G-coefficient</td>
<td>G-coefficient</td>
<td>Ordinal G-coefficient</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>0.829</td>
<td>0.903</td>
<td>0.790</td>
<td>0.900</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>0.835</td>
<td>0.900</td>
<td>0.814</td>
<td>0.899</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>0.870</td>
<td>0.901</td>
<td>0.812</td>
<td>0.899</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>0.886</td>
<td>0.900</td>
<td>0.837</td>
<td>0.900</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>0.893</td>
<td>0.901</td>
<td>0.849</td>
<td>0.901</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>0.897</td>
<td>0.900</td>
<td>0.859</td>
<td>0.902</td>
</tr>
</tbody>
</table>

Figure 3.2. Bias (%) of G-coefficients by number of scale points and item responses with no skewness for theoretical reliabilities of 0.90, 0.80, 0.60 and 0.40 for the crossed design
Figure 3.3. Bias (%) of the ordinal G-coefficients by number of scale points and item responses with no skewness for theoretical reliabilities of 0.90, 0.80, 0.60 and 0.40 for the crossed design.
Figure 3.4. Bias (%) of the G-coefficients by number of scale points and item responses with a negative level of skewness for theoretical reliabilities of 0.90, 0.80, 0.60 and 0.40 for the crossed design.
Figure 3.5. Bias (%) of the ordinal G-coefficients by number of scale points and item responses with a negative level of skewness for theoretical reliabilities of 0.90, 0.80, 0.60 and 0.40 for the crossed design.
Figure 3.6. Bias (%) of the G-coefficients by number of scale points and item responses with no skewness for theoretical reliabilities of 0.90, 0.80, 0.60 and 0.40 for the nested design.
Figure 3.7. Bias (%) of the ordinal G-coefficients by number of scale points and item responses with no skewness for theoretical reliabilities of 0.90, 0.80, 0.60 and 0.40 for the nested design.
Figure 3.8. Bias (%) of the G-coefficients by number of scale points and item responses with a negative level of skewness for theoretical reliabilities of 0.90, 0.80, 0.60 and 0.40 for the nested design
3.3.4 Discussion

The findings in this study resemble those reported by L. Muthén (1983) and Zumbo et al. (2007) for ordinal alpha in CTT. Compared to their traditional counterparts, the ordinal G-coefficients are consistently accurate estimates of the magnitude of the theoretical G-coefficients for all response categories and levels of skewness in the item response distributions. While the G-coefficients are all negatively biased, resulting in an underestimation of the theoretical G-coefficient, the ordinal G-coefficients are generally unbiased. The bias in the ordinal G-coefficients ranges from 0.0003% to 3%, while the underestimation of the G-coefficients range from 0.3% to 28%. The biased estimates are more prevalent for theoretical G-coefficients of 0.40
and binary response scales regardless of whether the data are treated as continuous or ordinal. Based on this simulation study, the G-coefficients provide a negatively biased estimate of the theoretical G-coefficients, unlike the ordinal G-coefficients, which exhibit very little, if any, bias.

When the data were normally distributed, the G-coefficient consistently led to the underestimation of the theoretical G-coefficient. On the other hand, the ordinal G-coefficient was consistently unbiased or slightly biased (generally positively). The difference in the magnitude of the bias between the G-coefficients and ordinal G-coefficients diminished as the number of response categories and the magnitude of the theoretical G-coefficient increased. This difference between the G-coefficients and ordinal G-coefficients was most pronounced when the number of response categories was less than five and the theoretical G-coefficient was poor (e.g., less than 0.60). In these situations, the ordinal G-coefficient more accurately represented the true reliability.

When the item response distribution was skewed, the difference in the magnitude between the G-coefficients and ordinal G-coefficients became more apparent. The G-coefficients were more biased in the presence of skewness. This bias, however, was less pronounced when the theoretical G-coefficient was 0.90 for a seven-point response scale. Under these conditions, the G-coefficient was 0.856 in the crossed design and 0.859 in the nested design, resulting in a 4.5% bias. On the other hand, the ordinal G-coefficient showed a slight positive bias or none at all. It was 0.900 in the crossed design and 0.902 in the nested design when the theoretical G-coefficient was 0.90. When the theoretical G-coefficient was greater than 0.40, the difference in value between the G-coefficients and ordinal G-coefficients was smaller, and it decreased even more as the number of response categories increased. Although the difference began to disappear as the number of response categories increased, it is still substantial even in the best-case
scenario (e.g., a theoretical G-coefficient of 0.90 with seven-point response categories). The G-coefficient was considerably affected by the skewness of the item distribution, but the ordinal G-coefficient was not. The results suggest that the ordinal G-coefficient should always be used when the item distributions are skewed for all response categories and magnitudes of reliability.

The difference between the G-coefficients and ordinal G-coefficients in the crossed and nested designs is subtle. For both coefficients, the systematic difference in bias for the nested design is much more variable than for the crossed design. In the crossed design for a binary response scale, the bias systematically lessens as the magnitude of the theoretical G-coefficient increases. However, this is not the case in the nested design, as shown in the binary, three- and four-point response scales for the different theoretical G-coefficients when items were normally distributed or skewed. This can be seen in the lines that cross over in Figure 3.6. In these instances, the systematic bias is not consistent. For example, the G-coefficient was more attenuated for the theoretical G-coefficient of 0.80 than 0.60 when the item distribution was skewed in the binary response option. This does not follow the pattern found in the crossed design, where the G-coefficient for a theoretical value of 0.60 is more attenuated than for 0.80. Perhaps one reason for this increased variability is the nature of nested designs, where the variances of the nested facets cannot be estimated and may lead to spurious results. In general, nested designs are not recommended for GT because not all sources of variance can be calculated. The unsystematic differences observed in the nested and crossed designs suggest that using the former may generate spurious, random or cautious values.

The design and skewness of the item response distributions exhibited an interesting relationship. In the crossed design, the ordinal G-coefficients were mostly underestimates of the theoretical G-coefficients when the item responses were normally distributed. In all other designs
and conditions, the ordinal G-coefficients were unbiased or slightly positively biased compared to the theoretical G-coefficients. The difference between the underestimation and overestimation of the ordinal G-coefficients in these particular cases may simply be an artifact of random variance. Although a positive bias can appear to be more dangerous than a negative bias, on average, it was never greater than 1%. The severest overestimation occurred when the response scale was binary. In this case, the bias was 3%, resulting in an ordinal G-coefficient of 0.412 compared to the theoretical value of 0.40. In the same circumstances, the G-coefficient underestimated the reliability by 19%, with a G-coefficient of 0.322. Therefore, the ordinal G-coefficient is a better estimate of the theoretical G-coefficient even if it leads to a slight positive bias. It is important to note that any variation of the ordinal G-coefficients above or below the theoretical G-coefficients occurred only at the third decimal place.

The greater accuracy of the ordinal G-coefficients can be attributed to the variance components estimation method and the modeling of the underlying latent variable. Considering the metric of the ordinal variables makes a substantial difference not only in the magnitude of the G-coefficient, but also in the magnitude of the variance components. In continuous CSM, the variance due to person and all its interaction terms was dramatically underestimated, while the error variance was consistently inflated, although these inaccuracies became less as the number of response categories increased. In contrast, the estimated variance components using ordinal CSM were consistently closer to the true values. The fact that the variance components are better estimates of the true values in ordinal than in continuous CSM explains why ordinal GT produces results that more closely resemble the true reliability than conventional GT.

These results have substantial implications for day-to-day research. Many researchers use GT to conduct D-studies, which allow them to optimize the reliability of the measurement tool
under development by changing the levels within each facet that contributes a significant amount of unwanted variance. A D-study is only as good as the estimated variance components and the computed G-coefficients. As one can see from this study, if ordinal data are treated as if it were continuous, it will result in the underestimation of the true reliability of the measure. More importantly, it will lead to an inaccurate estimate of the sources of error in the design, causing researchers to make wrong decisions about the number of items, raters or dimensions to include when conducting a D-study. Therefore, to make the most accurate choices in a D-study, one should use ordinal GT when the data are ordered categorical.

The results of this study suggest that even when the ordinal data are normally distributed, G-coefficients will consistently underestimate the theoretical value. This is even more apparent when the item distributions are skewed, which commonly occurs in ordinal data. Because it causes researchers to misinterpret the reliability of the measure under development, the underestimation of the G-coefficient has major ramifications, and may be responsible for the reconstruction or complete abandonment of their measure. Based on the results of this study, the pitfalls of treating ordinal data as continuous are too considerable to be ignored and warrant the use of ordinal GT whenever the data are ordinal. Ordinal GT will result in the most accurate representation of the true reliability of a measure and should therefore always be used when conducting a D-study.
Chapter 4: Discussion and Recommendations

Given that the findings for studies 1 and 2 have been discussed in Chapter 3, the purpose of this chapter is to summarize the results in a broader methodological context and to describe the novel contributions, limitations and recommendations with an eye toward future research directions.

4.1 Novel Contributions to the Field

This dissertation makes two broad novel contributions to the estimation of G-coefficients. The first is to provide a fuller explanation than was previously available of the model for estimating variance components to compute a relative G-coefficient for binary and polytomous response data via CSM. It created a proof of concept for estimating variance components via CSM for polychoric response data for both crossed and nested multi-facet assessment designs. It accomplished this by modeling the polychoric correlations of the underlying latent response variable. In a seminal though unpublished dissertation, L. Muthén (1983) estimated variance components and relative G-coefficients for a one-facet and three-facet fully crossed design using a variant of CSM techniques for binary response data. It is important to appreciate that L. Muthén was the first to demonstrate that variance components and a relative G-coefficient could be computed for binary responses by utilizing latent variable modeling techniques. Unfortunately, her work has gone widely unnoticed in the field.

This dissertation extends L. Muthén’s description of how and why CSM works as a method for estimating variance components for GT. Published studies describing the use of CSM (Marcoulides, 1996; Raykov & Marcoulides, 2006; Schoonen, 2005) simply cite earlier literature by Bock (1960), Bock and Bargmann (1966), Burt (1947) and Jöreskog (1978), but do not clearly explain why or how CSM results in the correct variance component estimations even
though not all factors of the design are explicitly stated in the model (e.g., the main effects of items and raters in a crossed design). This issue is important to highlight because in an ANOVA framework, all of the factors and their interactions are included, whereas in CSM, they are not. The truncated descriptions in the papers by Marcoulides, Raykov, and Schoonen are quite reasonable given their purpose, their audience, and their possible page limitations by the journals. Therefore, to encourage wider use by researchers and practitioners, a relatively modest part of the first novel contribution of this dissertation was to offer a psychometric and more accessible exposition of CSM as a method to estimate variance components in GT.

The second novel contribution is the quantification of the amount of bias in the G-coefficients when treating binary or ordinal data as continuous (i.e., ignoring the metric of ordinal variables) in estimating the relative G-coefficients. The underestimations ranged from 0.3% to 28%. The worst case is with skewed binary response data from a crossed design, wherein the estimated G-coefficient when treating the binary data as continuous is 0.285 when it should have been 0.40. In this worst-case scenario, using ordinal GT results in a G-coefficient of 0.399, thereby almost completely unbiased. This information about the amount of bias is invaluable to researchers because it provides practical implications of ignoring the metric of variables in computing reliability. The findings regarding the consistent downward bias found when treating the binary (and ordinal) item response data as if it were continuous suggests that these G-coefficients may be casting their assessment techniques in a negative light.

4.2 Implications for Researchers and Assessment Specialists

This dissertation reports on two studies. The first, the proof of concept, demonstrated that ordinal GT could be conceptualized using CSM of the tetrachoric or polychoric (depending on the number of response points in the item response distribution) correlation matrix for crossed or
nested designs. The second study compared ordinal GT to the more traditional approach, which treats ordinal data as continuous in estimating the reliability for different response categories and levels of skewness. The study found that when the metric of the ordinal variables is taken into account, the G-coefficient more accurately estimates the true reliability of a set of test scores.

The difference between ordinal and conventional GT is, to a varying extent, an effect of the number of scale points, the skewness, and the value of the theoretical G-coefficient. Of course, practitioners do not know the value of the theoretical G-coefficients, although they do control the other two factors. With this in mind, the difference between the two approaches is most pronounced when the item response distributions are skewed and five or fewer response categories are used. In these scenarios, one is clearly advised to use ordinal GT. However, when one has symmetric bell-shaped response distributions and six- or seven-point response categories, one could argue that the conventional G-coefficients are not dramatically different from ordinal G-coefficients and that there is not much to gain from using ordinal GT.

GT, or variants thereof such as coefficient alpha estimates of internal consistency, is used to decide whether an assessment can be applied in a high-stakes setting. For example, in medical education, an assessment tool is unacceptable unless it has a minimum reliability of 0.90. Assuming this is the true reliability of the assessment, a conventional G-coefficient of 0.90 would be obtained only if the item response data are symmetric bell-shaped and a seven-point response scale is used. This is the only instance in which conventional GT techniques (i.e., treating ordinal data as continuous) would correctly guide practitioners. In all other cases (e.g., response scale options or skewness of the response distributions), conventional GT techniques would never achieve a G-coefficient of 0.90 and the assessment would be incorrectly rejected. However, if ordinal GT were used in this situation, a value of 0.90 would be obtained regardless
of the number of response categories or whether the item response distribution was normal or skewed.

The results of this dissertation also have implications for D-studies. One purpose of conducting a G-study is to identify and estimate the variance due to each source of error. Then a D-study is performed to optimize reliability. The D-study therefore depends on the values obtained from the G-study. As a result, the D-study is only as good as the G-study’s variance components and G-coefficients. If the G-coefficients do not accurately represent the magnitude of each source of error, then the results of the D-study may not optimize the reliability of a measure. Since many D-studies are cost-benefit analyses, this can have serious implications. For example, researchers use D-studies to determine if they can achieve the same level of reliability if they employ one less rater or reduce the number of items on a survey to save time. The findings herein suggest that naively treating ordinal data as continuous causes the underestimation of the G-coefficient and the variance components. As a result, the D-study may reveal that one needs more raters, time points or items to obtain a higher reliability than what is actually required. To make the most accurate decisions and take advantage of what a D-study has to offer, practitioners should use ordinal GT when dealing with ordinal data. This has an important impact for researchers in education and psychology because G-coefficients are used to help optimize the measurement design. Those completing the questions on a measure can be burdened if incorrect decisions are made about the measurement design and this may also result in threats to the validity of the measure.

4.3 Limitations and Recommendations

The methods presented in this dissertation have two major limitations. First, only the relative G-coefficient can be estimated using the current method, parameterization and
conceptualization of the data because the covariance among the observed variables (e.g., items) is analyzed (R-method). One possible way to calculate an absolute G-coefficient and the variances due to the facet of generalization and its interactions is to structure the data so that the covariance among individuals is analyzed instead of the items (Q-method). Ancillary to this first limitation, only random facets can be modeled using CSM. This limitation impose a large constraint on the types of designs for which this method can be utilized in the estimation of G-coefficients. The results of this study suggest that discovering a solution to this limitation is paramount because of the benefits of ordinal GT.

The second limitation is that the findings in Study 2 of Chapter 3 are large-sample, population analogues. It was important to first determine if ordinal GT provides unbiased estimates (which imply population quantities) before studying its small-sample properties. Therefore, ordinal GT’s small-sample bias and variability when estimating G-coefficients is still unknown. A computer simulation study would need to be systematically conducted using the same experimental conditions manipulated in Study 2, with the addition of a sample size condition. It would be helpful to focus on the sampling variability and percentage of bias in the G-coefficient.

4.4 Conclusions

In light of the results of this dissertation, researchers are encouraged to avoid treating ordinal data as if it were continuous when computing relative G-coefficients. GT and the use of D-studies are advantageous when compared to conventional CTT; however, the benefits of GT may be reduced if the ordinal properties of the variables are ignored. Accurately representing the metric of the data will not only result in more reliable variance components and G-coefficients,
but in better D-studies. Ordinal GT results in the least amount of bias. For these reasons, ordinal GT should always be preferred over GT when dealing with ordinal data, especially given that statistical software is available to model ordinal data. If researchers use ordinal GT, they will discover that the grass is truly greener on the other side.
References


Appendix 1

The following is the LISREL syntax to conduct GT (CSM) for the two-facet, fully crossed design with person, four items and two occasions. The LISREL syntax to conduct GT (CSM) for a one-facet partially nested design is also provided, with person and items nested within occasions.

Crossed Design: MC

Observed Variables VAR1-VAR8

Raw Data from file data.LSF

Latent Variables person item1 item2 item3 item4 time1 time2

Relationships

'VAR1' = 1.00*person 1.00*item1 1.00*time1

'VAR2' = 1.00*person 1.00*item2 1.00*time1

'VAR3' = 1.00*person 1.00*item3 1.00*time1

'VAR4' = 1.00*person 1.00*item4 1.00*time1

'VAR5' = 1.00*person 1.00*item1 1.00*time2

'VAR6' = 1.00*person 1.00*item2 1.00*time2

'VAR7' = 1.00*person 1.00*item3 1.00*time2

'VAR8' = 1.00*person 1.00*item4 1.00*time2

Set the Covariances of item1 and person to 0.00

Set the Covariances of item2 and person to 0.00

Set the Covariances of item2 and item1 to 0.00

Set the Covariances of item3 and person to 0.00
Set the Covariances of item3 and item1 to 0.00
Set the Covariances of item3 and item2 to 0.00
Set the Covariances of item4 and person to 0.00
Set the Covariances of item4 and item1 to 0.00
Set the Covariances of item4 and item2 to 0.00
Set the Covariances of item4 and item3 to 0.00
Set the Covariances of time1 and person to 0.00
Set the Covariances of time1 and item1 to 0.00
Set the Covariances of time1 and item2 to 0.00
Set the Covariances of time1 and item3 to 0.00
Set the Covariances of time1 and item4 to 0.00
Set the Covariances of time2 and person to 0.00
Set the Covariances of time2 and item1 to 0.00
Set the Covariances of time2 and item2 to 0.00
Set the Covariances of time2 and item3 to 0.00
Set the Covariances of time2 and item4 to 0.00
Set the Covariances of time2 and time1 to 0.00

!Method of Estimation: Unweighted Least Squares
Path Diagram
End of Problem
Nested Design: MC

Observed variables VAR1-VAR12

Raw Data from file data.LSF

Latent Variables person r1 r2 r3

Relationships

'VAR1' = 1.00*person 1.00*r1
'VAR2' = 1.00*person 1.00*r1
'VAR3' = 1.00*person 1.00*r1
'VAR4' = 1.00*person 1.00*r1
'VAR5' = 1.00*person 1.00*r2
'VAR6' = 1.00*person 1.00*r2
'VAR7' = 1.00*person 1.00*r2
'VAR8' = 1.00*person 1.00*r2
'VAR9' = 1.00*person 1.00*r3
'VAR10' = 1.00*person 1.00*r3
'VAR11' = 1.00*person 1.00*r3
'VAR12' = 1.00*person 1.00*r3

Set the Covariances of r1 and person to 0.00
Set the Covariances of r2 and person to 0.00
Set the Covariances of r2 and r1 to 0.00
Set the Covariances of r3 and person to 0.00
Set the Covariances of r3 and r1 to 0.00
Set the Covariances of r3 and r2 to 0.00
Method of Estimation: Unweighted Least Squares

Path Diagram

End of Problem
Appendix 2

The following is the syntax to conduct ordinal GT (CSM) for a two-facet, fully crossed design with person, four items and two occasions. To conduct GT (CSM), just disable (by placing a “!” in front of) or remove the “CATEGORICAL ARE y1-y8” from the line of code under VARIABLE. When conducting a theta parameterization, the following estimators can only be used with ordinal data: WLS, WLSMV, and WLSM. This restriction is not due to statistical impossibility, but to the authors of Mplus.

TITLE: Ordinal GT two-facet fully crossed design with Mplus.

DATA: FILE IS data_file_name.dat;

VARIABLE:

   NAMES ARE y1-y8;
   USEVARIABLES ARE y1-y8;
   CATEGORICAL ARE y1-y8;

ANALYSIS:

   MODEL=nocovariances;
   PARAMETERIZATION=theta;
   ESTIMATOR=WLS;

MODEL:

   person BY y1-y8@1;
occasion1 BY y1-y4@1;
occasion2 BY y5-y8@1;
item1 BY y1@1 y5@1;
item2 BY y2@1 y6@1;
item3 BY y3@1 y7@1;
item4 BY y4@1 y8@1;
y1-y8;

OUTPUT: sampstat;
Appendix 3

The following are the thresholds for the symmetric and -2 skewed Likert response categories. As a demonstration of how to read this table, for a five-point response format that results in a symmetric item response distribution, there are four thresholds: -1.8, -0.6, 0.6, and 1.8.

Thresholds for symmetric Likert response category distribution

<table>
<thead>
<tr>
<th>Thresholds</th>
<th>Likert Response categories</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1.5</td>
</tr>
<tr>
<td>4</td>
<td>1.8</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>2.14286</td>
</tr>
</tbody>
</table>

Thresholds for -2 skewed Likert response category distribution

<table>
<thead>
<tr>
<th>Thresholds</th>
<th>Likert Response categories</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>1.062519302</td>
</tr>
<tr>
<td>2</td>
<td>1.298836633</td>
</tr>
<tr>
<td>3</td>
<td>1.2816</td>
</tr>
<tr>
<td>4</td>
<td>1.6546</td>
</tr>
<tr>
<td>5</td>
<td>1.8002</td>
</tr>
<tr>
<td>6</td>
<td>2.1002</td>
</tr>
</tbody>
</table>
Appendix 4

The following are the latent regression models for the crossed and nested designs used to generate the simulated data in Study 2.

The crossed design, $Y_{jk}$, consists of four items ($j = 1, 2, 3, 4$) and two raters ($k = 1, 2$) randomly selected:

\[
Y_{1j} = b_{1j} \eta_p + b_{2j} \eta_{p1} + b_{3j} \eta_{p2} + e_{1j} \\
Y_{2j} = b_{4j} \eta_p + b_{5j} \eta_{p1} + b_{6j} \eta_{p2} + e_{2j} \\
Y_{3j} = b_{7j} \eta_p + b_{8j} \eta_{p1} + b_{9j} \eta_{p2} + e_{3j} \\
Y_{4j} = b_{10j} \eta_p + b_{11j} \eta_{p1} + b_{12j} \eta_{p2} + e_{4j} \\
Y_{1k} = b_{13k} \eta_p + b_{14k} \eta_{p1} + b_{15k} \eta_{p2} + e_{1k} \\
Y_{2k} = b_{16k} \eta_p + b_{17k} \eta_{p1} + b_{18k} \eta_{p2} + e_{2k} \\
Y_{3k} = b_{19k} \eta_p + b_{20k} \eta_{p1} + b_{21k} \eta_{p2} + e_{3k} \\
Y_{4k} = b_{22k} \eta_p + b_{23k} \eta_{p1} + b_{24k} \eta_{p2} + e_{4k}
\]

The nested design, $Y_{jkl}$, includes two randomly selected items ($j = 1, 2$) on three separate occasions ($k = 1, 2, 3$) by six randomly selected raters ($l = 1, \ldots, 6$):

\[
Y_{11l} = b_{11l} \eta_p + b_{12l} \eta_{p1} + b_{13l} \eta_{p2} + e_{11l} \\
Y_{21l} = b_{14l} \eta_p + b_{15l} \eta_{p1} + b_{16l} \eta_{p2} + e_{21l} \\
Y_{12l} = b_{17l} \eta_p + b_{18l} \eta_{p1} + b_{19l} \eta_{p2} + e_{12l} \\
Y_{22l} = b_{20l} \eta_p + b_{21l} \eta_{p1} + b_{22l} \eta_{p2} + e_{22l} \\
Y_{13l} = b_{23l} \eta_p + b_{24l} \eta_{p1} + b_{25l} \eta_{p2} + e_{13l} \\
Y_{23l} = b_{26l} \eta_p + b_{27l} \eta_{p1} + b_{28l} \eta_{p2} + e_{23l} \\
Y_{14l} = b_{29l} \eta_p + b_{30l} \eta_{p1} + b_{31l} \eta_{p2} + e_{14l} \\
Y_{24l} = b_{32l} \eta_p + b_{33l} \eta_{p1} + b_{34l} \eta_{p2} + e_{24l} \\
Y_{15l} = b_{35l} \eta_p + b_{36l} \eta_{p1} + b_{37l} \eta_{p2} + e_{15l} \\
Y_{25l} = b_{38l} \eta_p + b_{39l} \eta_{p1} + b_{40l} \eta_{p2} + e_{25l} \\
Y_{16l} = b_{41l} \eta_p + b_{42l} \eta_{p1} + b_{43l} \eta_{p2} + e_{16l} \\
Y_{26l} = b_{44l} \eta_p + b_{45l} \eta_{p1} + b_{46l} \eta_{p2} + e_{26l}
\]