UNDERSTANDING THE BULLYING STRUCTURE ON A COMPLEX NATIONAL DATA SET AND SAMPLE SIZE REQUIREMENT FOR NON-NORMAL COMPLEX MULTILEVEL DATA FOR THE MULTILEVEL STRUCTURAL EQUATION MODEL

by

ADEYA SHONTELLE POWELL

(Under the Direction of Martha Carr)

ABSTRACT

This study consisted of two parts. The first study sought to explore the relationship between bullying and depression/suicide using the most frequently used survey given to American youths, the Youth Risk Behavior Survey (YRBS), and to employ one of the least used statistical tools for this survey, Structural Equation Modeling. Using a mediated structural model with two possible mediators, school violence and teen alcohol abuse and use, school violence was shown to mediate the relationship between the dichotomous bullying variable and the depression/suicidality factor, while teen alcohol abuse and use was not a mediator. The second part of the study analyzed sample size requirements for a multilevel structural equation model when non-normality was present. This study had two subparts: non-normal continuous and non-normal non-continuous/categorical sample size requirement. Robust Maximum Likelihood estimator (MLR) was the only estimator to perform well at both the between and within level on non-normal continuous data. Weighted Least Square estimators did not perform well on categorical data when the sample size was 100 or less.

INDEX WORDS: MSEM, multilevel, YRBS, CDC, bullying, school safety, suicide, SEM

UNDERSTANDING THE BULLYING STRUCTURE ON A COMPLEX NATIONAL DATA SET AND SAMPLE SIZE REQUIREMENT FOR NON-NORMAL COMPLEX MULTILEVEL DATA FOR THE MULTILEVEL STRUCTURAL EQUATION MODEL

by

ADEYA SHONTELLE POWELL

B.S., Georgia State University, Mathematics 2005

M.A., University of Georgia, Mathematics Education 2006

M.S., University of Georgia, Statistics 2012

A Dissertation Submitted to the Graduate Faculty of The
University of Georgia in Partial Fulfillment of the Requirement for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2013

UNDERSTANDING THE BULLYING STRUCTURE ON A COMPLEX NATIONAL DATA SET AND SAMPLE SIZE REQUIREMENT FOR NON-NORMAL COMPLEX MULTILEVEL DATA FOR THE MULTILEVEL STRUCTURAL EQUATION MODEL

by

ADEYA SHONTELLE POWELL

Approved:

Major Professor:	Martha Carr
Committee:	Debaroh Bandalos
	Stephen Olejnik
	William McCormick

Electronic Version Approved:

Dr. Maureen Grasso
Dean of the Graduate School
The Univerity of Georgia
December 2013

ACKNOWLEDGEMENTS

This dissertation could not be done without the help of Dr. Debbie Bandalos, Dr. Marty Carr, and Dr. Steve Olejnik. I am very grateful for their tireless help, frankness, and guidance. Thanks to Dr. Tiffiany Ahoulou for her encouragement and sisterhood. Thanks to the Southern Regional Educational Board (SREB) and their director, Dr. Abraham, for being more than a check and a handshake.

TABLE OF CONTENTS

Page

SECTION

CHAPTER

SECTION

CHAPTER

LIST OF TABLES

LIST OF FIGURES

Page

SECTION ONE

UNDERSTANDING THE BULLYING STRUCTURE ON A COMPLEX NATIONAL DATA SET

STUDY ONE INTRODUCTION

Bullying has attracted national media attention as a result of its links to teen suicides and episodes of school violence, such as the student-lead massacre at Columbine High-School in April 1999. Society's increased concern for school safety has attracted the attention of lawmakers and educators (Larkin, 2007). Prior to the Columbine shooting, the National Educational Goals Panel of 1993 had already recognized the problem of school violence, and stated as one of its goals that "by the year 2000, every school in America will be free of drugs and violence and will offer a disciplined environment that is conducive to learning" (as cited in Batche, 1994). President Clinton included these goals in the Guns Free Schools Act of 1994 (Skiba, 2000), which created a federal law to expel students for up to a year for bringing weapons to schools. After the school shooting at Columbine High-School, state lawmakers expanded the Guns-Free School Act to include various infractions such as fighting, swearing, and disrupting class (Skiba, 2010).

To enact President Clinton's amendment, America's war on drugs policy, from the 1980s, was adapted for the school environment as a way to control students (Skiba & Peterson, 1999). Security measures including metal detectors, maintaining locked school doors at all times, having a detail of security staff, as well as automatic expulsion of students for violations of school safety rules, created what some call a prison-like environment (Skiba & Peterson, 1999). This approach to dealing with school violence backfired. Mayer and Leone (1999) found that a policed school environment, which involved the use of metal detectors and on-site police officers, led to more disruption and violence at schools. Conversely, Skiba & Peterson (1999, p.8) found that “schools with no reported crime were less likely to have a zero tolerance policy (74%) than schools that reported incidents of serious crime (85%).” Just as the war on drugs failed to solve society’s ills, the Guns Free School Act did little to solve the problem of school violence or create a positive school climate (Gray, 2001).

The failure of the American school system is contrasted with a successful program in Norway. In 1983, Norway faced a similar challenge of teen bullying when three students who were targets of violent bullying committed suicide. Enlisting the help of Norwegian researcher and professor, Dan Olweus, the Norwegian Ministry of Education initiated a nation-wide campaign against bullying that later became known as the Olweus Bullying Prevention Program (OBPP). Olweus’ scientific inquiry into bullying provided the academic world a refined definition of bullying and strategies on how to effectively reduce bullying.

Schools employing Olweus’ methodology were able to reduce bullying by up to 50% within a two year period; his procedure became the world’s most widely emulated model (Kalman, 2011). The impact of the program depended on the grade level with lower grades, such as middle or elementary, taking less than a year to achieve a significant reduction. High schools took longer to show the impact of the program. The OBPP program takes a holistic multilevel multidimensional approach. Olweus (1994) concluded that there was a need to restructure the entire social environment at the school, class, and individual levels. Specifically, Olweus changed the school climate to build a sense of community among students and adults (Olweus, 2012).

School climate defines the quality and character of the school by focusing on four essential concepts: safety (school norms and rules, physical and emotional safety); relationships (respect for diversity, school connectedness/engagement, social support for adults/teachers, social support for students, leadership); teaching and learning (social, emotional, and ethical learning, support for learning, professional relationships); and institutional environment (physical surrounding) (School Climate Research Summary, 2010). Focusing on these four basic concepts allows for a safe school environment; which in turn fosters school attachment and support learning (Center for Social and Emotional Education, 2010). Setting school rules against bullying that are known and modeled by all the key stakeholders (e.g. teachers, staff, and students) is the glue of a positive school environment. Osterman (2000) noted that conditions in the classroom and school influence students' feelings about themselves, which are reflected in student engagement and achievement. Blum, McNeely, & Rinehart (2002) concluded that when adolescents who felt cared for by their school and considered themselves to be a part of their school were less likely to use substances, engage in violence, or initiate sexual activity at an early age. By taking a more comprehensive evidenced based approach to bullying, Olweus' blueprint program reduced both bullying and anti-social behaviors such as vandalism (Clemson.edu/Olweus, 2012). This work showed the importance of school climate for bullying.

Suicide and Bullying

There have been only two comprehensive published literature reviews on the research connecting bullying and suicide within the past five years. Kim and Leventhal (2008) did a systematic literature review of studies on bullying and suicide using six databases: Web of Science, SCOPUS, EMBASE, PubMed, PsychInfo, and Ovid Medline. Of the 867 papers found to contain some bullying reference, only 103 addressed the relationship between bullying and suicide. Of the 103 papers found, only 37 specifically targeted teens, employed a quantitative component, and provided clear descriptions of measures for bullying or suicidal behavior. In the summary of their findings, 92% used self-reported bullying while only 5% (two studies) used peer nomination to identify bullying; nearly all the of the studies used odds ratios (OR) as their statistical method; all 37 were cross-sectional surveys; 73% ignored special populations such as homosexuals, bisexuals, or populations with developmental disorders; only about half of the studies were conducted in the U.S. (one third was conducted in Europe); and of the 17 US studies, over half (nine) used data from the Youth Risk Behavior Survey (YRBS; Kim & Leventhal, 2008)

Kim and Leventhal (2008) reported that most of the studies reported positive associations between suicidality (thoughts, plans, and attempts) and bullying. However, there are many methodological short comings in the research. Most of the studies failed to control for known suicide risk factors such as depression, history of suicide, substance abuse, or emotional distress, and most studies relied heavily on self-reports of bullying. Consequently, there might be some misinterpretation based on underlying issues. Lastly, each study was cross sectional which makes causal inferences impossible.

In response to these shortcomings, Kim and Leventhal (2008) suggested that future studies should address the issue of a causal relationship between bullying and suicidality (such a study would be impossible), have multiple informants for identifying the victims of bullying, include more known confounding effects, and should be more inclusive of special populations. They noted only one study controlled for gender, past suicide attempts, and depression. Relatedly, Klomek et al. (2011) noted that cross-sectional studies have a problem with generalizability. Specifically, most suffer from shared method variance caused by having the same person identify both who is bullied and if he or she has any suicidal risks. Furthermore, most assessments use only brief screening instruments (few items) to assess suicide related thoughts or behavior, and not all studies include a definition of bullying. Lastly, most only assessed the association between bullying and suicidal ideation, with a few assessing the relationship between bullying and suicidal behavior. These associations are unable to provide adequate evidence that bullying does more than merely correlate with suicide behavior. The current study will try to provide more understanding between bullying and suicide by seeing if alcohol is a contributing force and if the structure is the same or different depending on gender.

The Current Study

The current study examines the relationship between bullying, suicidal thoughts, school climate (violence), and alcohol abuse from a new perspective. Although SEM is still correlational, this study tested various structural models regarding what relationship might be present, which one is stronger, and whether gender differences existed.

Kim and Leventhal (2008) found over half of the studies on bullying in the U.S. were done using the Youth Risk Behavioral Survey (YRBS). YRBS is a national survey given every two years to monitor risk behavior in teens which provides longitudinal data on bullying, alcohol and substance abuse, suicidal thoughts, and violence in schools. The YRBS is a self-report survey that covers six broad risk behaviors: injury and violence, alcohol and drug use, tobacco use, nutrition, physical activity, and sexual behavior. The YRBS survey does change from year to year, but 2009 marked the first time a direct yes or no question was asked about bullying; the 2011 survey included a question on cyber bullying. Prior to 2009, all the questions primarily inquired about school safety or violence. The surveys dealt with more general school violence but included a few questions on bullying such as not going to school because the student felt unsafe at school or being threatened or injured with a weapon at school (Kim, 2008).

The current study used structural equation modeling to analyze data of interest from the YRBS data set. The Center for Disease Control (CDC) recommends the use of logistic regression when analyzing the YRBS data set. Structural equation modeling (SEM) is better in that you can model multiple questions at one-time instead of separate equations. This accommodates cases where variables are both outcomes and covariates, allows for the inclusion of variables that are closely related, and allows for unobserved variables to be incorporated in your model. Kupek's (2006) study found SEM classification was similar to logistic regression but with more flexibility.

One of the most important advantages of SEM is the correction of measurement error in the model. This is done by defining more than one item or survey question to describe a latent or unobserved construct (Bedelian et al., 1997). Another advantage is the ability to test various causal assumptions by comparing competing models that are all plausible based on theory. Rather than advocating a causal claim, SEM is more probabilistic in that the various plausible models can be compared and ruled out. Lastly, SEM is very well suited for doing mediated analysis. Using a regression based approach on mediation analysis leads to inflated direct effects and attenuated mediated effects (Baron & Kenny, 1986; as cited in Fabrigar et al., 2010).Therefore, although SEM is a correlational technique and is not inherently better than any other technique when it comes to non- experimental data, it is useful when measurement error is present and for mediation analysis. Because measurement error is frequently involved in survey data, I decided SEM would be a great tool to do my current research. It might help me detect relationships that might not show up using standard statistical tools.

YRBS is one of the most common surveys used on US children but no one has studied the risk of suicide in relationship to bullying using the national data set. My study seeks to analyze the relationship between bullying and depression/suicidality using mediation analysis within the context of SEM. Specifically, does school violence mediate the relationship between the two variables, or is alcohol a mediator (or both)? The specific research questions to be addressed in this study are as follows:

1) Is there a relationship between bullying and depression/suicidality?

2) Is that relationship mediated by school violence?

3) Is teen alcohol abuse and use another possible mediator in the model?

I hypothesize that there will be a relationship between bullying and depression/suicidality; that this relationship will be mediated by school violence; and that alcohol abuse and use will be a competing mediator that explains the relationship between bullying and depression/suicidality. Teen alcohol abuse and use was included as a possible competing mediator because alcohol interacts with depression and life stressors to contribute to suicide, which is the third leading cause of death for people ages 14 to 25 according to the Center for Disease Control (CDC).

CHAPTER TWO

STUDY ONE LITERATURE REVIEW

This chapter includes a literature review of the YRBS instrument and some of the variables measured by YRBS that will be examined in this study including, bullying, alcohol use and abuse, and teen suicide. Several definitions for alcohol use and abuse and teen suicide are mentioned but no one definition is chosen so as to maximize the amount of construct related questions included in the present analysis. After the introduction to the variables included in this study, a brief introduction to SEM is presented followed by an introduction to Mediation Modeling; additional details are provided about the benefits and short comings of SEM.

The YRBS instrument

The Youth Risk Behavior Survey (YRBS), first administered in October 1989, is a multifaceted survey that was designed to meet the need of national health initiatives and state and local governments to monitor health-risk behaviors that contribute to the "leading causes of death, disability, and social problems among youth" (CDC, 2004, p.1). Developed by the Center of Disease Control (CDC), the YRBS does this by monitoring risk behavior over time and looking at trends. The information gathered by YRBS allows state and local partnerships to evaluate the effectiveness of their initiatives, and assess if the national health initiatives are being met. Prior to its development in the late 80s, there were two health surveys that had been used to gather health information about teens: *Monitoring the Future: A Continuing Study of the Lifestyles and Values of Youth*, which measured drug use in 12th graders, and *the National Adolescent Student Health Survey*, 1987, which was a one-time survey of teen risk behaviors from grades 8th-10th (CDC, 2004). Both surveys were either limited in scope or were not administered on an on-going basis. Because these limitations did not serve the needs of state and local governments, the CDC designed the YRBS survey. YRBS allows state and local governments the flexibility to modify the questions on YRBS for a different target population or purpose. Geared primarily for 9th-12th graders, some state and local governments have modified YRBS, to address risk behavior for alternative school children, middle school children, and "special populations" such as American Indian youth (CDC, 2004). This partnership between federal and state agencies is facilitated by grants or cooperative agreements.

The Youth Risk Behavior Surveillance System (YRBSS) monitors six categories of youth health-risks: "behaviors that contribute to unintentional injuries and violence; tobacco use; alcohol and other drug use; sexual behaviors that contribute to unintended pregnancy and sexually transmitted diseases (STDs); unhealthy dietary behaviors; and physical inactivity" (CDC, 2004, p.3). YRBSS does this monitoring through the use of the YRBS Instrument. For example, in 1993, the *National Educational Goal* was for a safe and drug-free school, so, questions were added to reflect those national goals (CDC, 2004). In 1999, due to the obesity epidemic, questions about the body mass index (BMI) were added to the survey. Recently, the 2011 YRBS added another question on bullying because of the link between bullying and depression in teens. Since 1991, the YRBS has been given every odd numbered year on a biennial basis. Participation is voluntary and parental permission must be obtained prior to participation. Students' identifying information is not attached to a particular survey, ensuring anonymity. The school response rate was 81%, the student response rate was 88%, with an overall response rate of (.81) *(.88)= 71%.

The national version of YRBS uses a three stage complex sampling design. During the first stage of the design, about 1200 large-sized counties or groups of "smaller adjacent" counties are selected (CDC, 2009, p.2). From these, counties are subdivided into 16 non-overlapping groups or strata based on their density and ratio of minorities. Roughly 57 counties are then selected with school enrollment size as the criterion of selection. During the second stage, almost 200 schools, public and private having grades 9th-12th, are selected from the 57 counties based on their school enrollment size. Schools with a higher enrollment of minorities are sampled at a higher rate. The third stage consists of a random sample of one or two required classes from grade 9th-12th within the selected school. The entire class is selected to be in the study, students have the option to not participate.

To adjust for oversampling minorities, non-responses, and the unequal probability of selection, YRBS employs a weighting of each student to keep the sample more representative of the target population, 9th to 12th grade students.

Every biannual survey does vary slightly in regards to the number of questions being asked and content in accordance with the current national objectives. Most questions are ordinal in scale, however, there are a small number of dichotomous variables. Most of the ordinal variables have at least five progressive categories. Along with the health-risk questions, demographic background information is collected such as gender, age, grade level, race, and geographic region.

Validity and Reliability of the YRBS

In general, if we are measuring someone’s weight, the scale is reliable if the same number appears again and again, but it might not be valid. For a test to be valid it must be reliable, but there is more to validity than reliability. Validity involves inferences drawn about the test, more accurately, is the inferences drawn supported by the test scores. For example, the weight-scale might reliably give you a measure of a person’s weight but can someone now conclude that this is someone's true weight? For example, a non-calibrated scale might actually produce weight measurement above or below a person's actual weight. Without the calibration, the scale might be consistent but not accurate. Validity measures the degree of accuracy and reliability measures consistency.

Test-retest reliability tests for YRBS were done on the national level data in 1992 and in 2000. On both occasions, a convenience sample of 7th to 12th grade students were administered the test as scheduled (i.e., 1991 and 1999) and approximately a fortnight later a second administration occurred. The 1991 test-retest reliability evaluation, showed the overall level of agreement was "substantial or high" for seventy-five percent of the items (CDC'04). The reliability or level of agreement did not fare as well for younger, 7th grade students. Using Cohen's kappa, as a measure of agreement between categorical items, the 2000 test-retest reliability evaluation deleted ten items with low reliability, namely, items about personal injury during physical activities (CDC'04). Items related to school climate, such as, carried weapons to school, felt unsafe while at school, and fighting on school property, had a moderate degree of reliability – the kappa was between .41 and .60 (Brener, 2002). Items relating to substance abuse, such as alcohol use and marijuana use, had a high level of agreement—kappa was between .61 and .80 (Brener, 2002). Items relating to depressed mood or depression also had a high level of agreement between both administrations (Brener, 2002). The national 2009 YRBS survey does not include younger students.

No formal validity study has been conducted on the YRBS but individual questions have been selected and studied. In 2000, after completing the second round of YRBS, students were physically measured for weight, height, and BMI. The results of the physical measure were compared with their self-reported questions on their height and weight. The students reliably gave the same answers to repeated administrations of the questionnaire, but they actually over-reported their height and under-reported their weight. Inferences drawn about their height and weight would have been inaccurate. Brener (2003) found that cognitive and situational influences especially hurt the validity of self-reported health-risk surveys. Cognitive factors are internal features that entail the mental capacity to recall, make decisions, or comprehend. Situational factors are factors that cause response bias, such as, fear of judgment and anxiety about anonymity. Furthermore, Brener (2003) found that substance abusers had a difficult time recalling their use of substances, especially after an extended amount of time, and substance abuse during the administration of the survey compounded the inaccuracies. Better recall was found for short question stems and short test duration. Lastly, the type of administration and gender affected whether or not adolescents were truthful about substance abuse. Specifically, pencil and pen self-administered surveys were more accurate and some researchers found that validity on these surveys depends on gender (Brener, 2003).

A small state level validity study on suicidal ideation items on the YRBS was conducted in 2010 to measure the convergent and discriminant validity. Convergent validity means that two theoretically closely related constructs are found to be highly correlated; discriminant validity means two theoretically dissimilar ideas are shown to have a low correlation, such as happiness and hypochondria. Klonsky (2010) compared the Patient Health Questionnaire for Adolescents (PHQ-A), the McLean Screening Instrument for Borderline Personality Disorder (MSI-BPD), and the UCLA Loneliness Scale (UCLA) to analyze the middle-school and high-school YRBS suicide items. There were three questions relating to suicide on YRBS: 1) Have you ever thought about killing yourself? 2) Have you ever made a plan about how you would kill yourself? 3) How many times have you actually tried to kill yourself? Klonsky (2010) defined lifetime suicidal ideation as the thought or planning of suicide, excluding attempts. Both suicidal ideation and attempts were found to be related to items on depression, anxiety, and loneliness. YRBS ideation items tended to be more strongly associated to other items on ideation than attempts. Items on attempts were strongly correlated with items on MSI-BPD corresponding to self-harm, whereas ideation items were not as strongly correlated (Klonsky & May, 2010). This demonstrates some validity for the items on suicide ideation and attempt.

Empirical Studies using YRBS

During an online search, using Google Scholar, a literature review of research using YRBS data found logistic regression was the primary statistical analysis employed. For example, Mays (2009), using six logistic regression models on the YRBS 2005 data, found male athletes were more likely to be involved in heavy drinking and drunk driving than females. Miller (2007), using logistic regression on the 2003 National YRBS, found that binge drinking increased with age and grade, and that it was associated with other risk-factors such as smoking, attempting suicide, and using illicit drugs. Cavazos-Rehg (2010), using multinomial logistic regression, found the more risk factors a teen had the greater chance of that teen having multiple teen pregnancies. Martins (2008) demonstrated a relationship between drug and alcohol use (i.e., ecstasy, marijuana, alcohol and tobacco use) and moderate to low academic achievement, using multinomial logistic regression. Using logistic regression on YRBS 2005 data, Epstein (2009) was able to show alcohol/drug use, aggression, risky sexual behaviors, and health problems were all associated with suicide. Jiang (2010), using multivariate logistic regression, found that feeling safe at school, sexual orientation, and immigration status had a high association with adolescent suicide. Behnken (2009) found, using logistic regression and mediational analysis, binge drinking explains the relationship between suicide and sexual victimization. Using cluster and multivariate logistic regression, Paxton (2007) found that those that engaged in health risk behaviors such as binge drinking, drug use, and sexual risk behavior were more likely to report having depressed moods.

In regards to SEM, Fabrigar and colleagues (2010), saw causal inferences as a continuum, from a single casual interpretation to an all causal assumption. They viewed SEM as having no more of a basis to infer causal conclusion than any other statistical technique (i.e., SEM cannot make up for a non-experimental study). However, SEM can be used to evaluate competing causal claims.

Bullying literature review

What is bullying? When does bantering turn into something more serious? Is bullying something that can be observed? Bullying has been described as a form of violence and aggression; however, many people still have a hazy concept of what constitutes bullying because bullying manifests in various observable and non-observable forms (Orpinas & Horne, 2006). Horne (2003) describes bullying as being on a continuum between playful behavior and delinquent criminal acts. According to Newman (2000), bullying is a form of aggression that involves intent, imbalance of power, and repeated acts of aggression (Neman & Horne, 2000).

Usual observers, such as teachers, cannot see intent so they have a hard time distinguishing when behavior has crossed from playfulness to something a little more serious (Orpinas & Horne, 2006). Orpinas (2006) believes most people have experienced some form of aggression in their life but bullying is repeated aggression (Orpinas & Horne, 2006). “Repeated acts of aggression generate a deeper level of fear and intimidation than an isolated event”

(Orpinas & Horne, 2006, p.15). Spreading rumors, sexual harassment, or physical assaults are a few of some of the myriad forms of bullying. Since bullying involves imbalance in power, the various imbalance categories can serve as an indicator of who might be the bully and who might be the victim. Imbalances can take the form of inequities in physical attributes such as weight or height, inequities in intellectual abilities, or inequities in social skills (Orpinas &Horne, 2006). One way aggression is based on the type of act such as physical, verbal, relational, and sexual acts that cause psychological or physical harm to others (Orpinas & Horne, 2006, p.24). The various types of aggression can be determined by the harm caused.

Consequences of bullying. According to Hazler (1992), almost 76% of students have experienced incidents of bullying while 14% are frequent victims of bullying (as cited in Newman & Horne, 2000). Studies have shown constant victimization from bullying results in psychological or psychosomatic disorders such as bedwetting, headaches, or depression (Orpinas & Horne, 2006). Research has found some forms of these disorders continue into adulthood in the form of anxiety, adult intimacy issues, or depression (Gilmartin, 1987; Gladstone et al., 2006, Olweus, 1997). Since bullied school children are more likely have anxiety and fear about school, they often avoid school and can become withdrawn (Batsche & Knoff, 1994; Berthold & Hoover, 2000). Research shows these children are very likely to drop out of school, have poor grades, and show an increase in aggressive behavior such as bringing weapons to school, and possibly becoming a bully themselves (Baker et al., 2008; Batsche & Knoff, 1994).

Those who bully are more likely than non-bullies to smoke, drink, or bring weapons to school (Berthold & Hoover, 2000; Nansel et al., 2003). Bullying behavior as a child serves as precursor to future delinquency. Child bullies were likely to demonstrate antisocial behavior in the form of alcohol use, property crimes, skipping school, and bringing weapons to school

(Berthold & Hoover, 2000; Liang et al., 2007). According to a longitudinal study done by Olweus and Alsaker (1991), as bullies transitioned to adulthood, criminality and law enforcement issues were prevalent (as cited in Berthold & Hoover, 2000). “Approximately 60% of boys who were characterized as bullies in grades 6-9 had been convicted of at least one officially registered crime by the age of 24. Even more dramatically, as much as 35-40% of the former bullies had three or more convictions by this age…” (Olweus, 1997, p. 501).

Bullies can also be victims of bullying. The intersection of bully and victim creates a third category called bully-victims. Studies have shown that both bullies and victims are more likely to suffer from depression than those who have never been involved in any bullying incident (Seals & Young, 2003). Just like the line between bully and victim is not necessarily independent in terms of mental outcomes, the line between bully and victim is not necessarily independent in definition. Bully-Victims are sometimes bullies and sometimes victims. Bully-Victims are three times more likely to report being a victim of bullying; have more depressive symptoms; and have more behavioral problems (Haynie et al., 2001). Stein’s (2006) research on male bullies, victims, and bully-victims, also found that bully-victims suffered more psychologically and had more delinquency issues than did those who were solely bully or solely victims. Although bully-victims are rare, more research has to be done on bully-victims.

School climate and bullying. Because the environment where bully and victims often cross paths is the school or classroom, school becomes a critical juncture or crossroad, and school climate moderates the path taken for both the victim and bully. Since bullied school children are more likely to have anxiety and fear about school, they often avoid school and can become withdrawn (Batsche & Knoff,1994; Berthold & Hoover, 2000). Research shows these children are very likely to drop out of school, have poor grades, and according to some research, show an increase in aggressive behavior such as bringing weapons to school, and possibly becoming a bully themselves (Baker et al., 2008; Batsche & Knoff,1994*)*. Bullying behavior as a child not only serves as precursor to future delinquency but research has found that child bullies were likely to demonstrate antisocial behaviors as children in the form of alcohol use, property crimes, skipping school, and bringing weapons to school (Berthold & Hoover, 2000; Liang et al., 2007)

As a way of addressing a growing threat to public health, many schools adopted a zero tolerance policy to improve school safety, but research has shown those policies were too premature and did not have the intended effect. In fact, the policies lead to blanket punitive actions, such as suspensions, which lead to an increase in punitive actions rather than an increase in understanding about the nature or structure of bullying (Orpinas &Horne, 2006; Sampson, 2002). Mayer and Leone (1999) found that a policed school environment, such as the use of metal detectors and on-site police officers, led to more school disruption and violence at schools. Thus, zero-tolerance policies and policed schools do not decrease bullying. Peer mediation, group therapy, and advocating assertiveness were also failed strategies that were deemed ineffective in addressing bullying (Sampson, 2002). Each of those strategies cast the victim as an equal participant who has control over what was happening to them. For example, these interventions often require the victim to stand up for him or herself or confront the bully through peer mediation. They may also focus on improving the victim's self-esteem with group therapy. These strategies have been proven to be ineffective.

Orpinas (2006) describes a climate as an environment that can either bring out the best or worst qualities. A positive school climate is a multidimensional construct that encompasses the physical aesthetics of the school, interpersonal interactions between teachers and students, interpersonal interactions between students, and the school policies (Orpinas & Horne, 2006, p.80). Orpinas and Horne (2006) believe that there are eight factors that contribute to a positive school climate, namely, excellence in teaching, school values, awareness of strengths and problems, policies and accountability, caring and respect, positive expectations, teacher support, and physical environment.

Excellence in teaching encompasses teaching ability, classroom management skills, and motivational skills (Orpinas & Horne, 2006). Hein (2004) found that good teaching and appropriate methods to motivate students reduces behavior problems (as cited in Orpinas & Horne, 2006). Because bullies often have academic problems, a positive school environment is one where the teacher is aware of the students' abilities and needs, and can address them (Orpinas & Horne, 2006). Classroom management skills are also critical because the teacher must address disrespectful behavior and bullying (Orpinas & Horne, 2006). Teachers and administrators are instrumental in setting the tone for expected or acceptable behavior. In fact, since bullying can occur in various school locations (e.g. restrooms or hallways), having all the stakeholders understand what is acceptable behavior leads to a positive school environment (Orpinas & Horne, 2006). Having clear and structured set of acceptable behavior rules, which are implemented with consequences, has been found more effective than a restrictive, zero tolerance, prison-like school environment (Mayer & Leone, 1999). Self-protection, such as bringing guns to school, is seen less frequently in environments that focus more on internal controls than external behavioral controls (Mayer & Leone, 1999). Even when after controlling for the lack of involvement of a critical stakeholder, the parents, or other risk factors (e.g. poor peer relations), Espelage and Swearer (2009) found an effective anti-bullying school environment serves as a buffer (as cited in Swearer et al., 2010).

YRBS and bullying. The YRBS has mainly focused on questions about weapons carrying, fighting or violence, suicide, substance abuse, weight and weight maintenance issues, and sexual behavior. The 2009 national version of YRBS was the first time students were asked directly about being bullied at school. In the 2011 version, a second question on bullying was added that asked about electronic bullying through social network sites and email. The YRBS has some indirect questions about the school environment that can be considered indicators of bullying or a hostile school climate. Specifically, the YRBS asks about the number of days a student has felt unsafe at school and whether he/she has been threatened or injured at school.

Comparatively, little research has been done with the YRBS instrument where the focus was bullying. When YRBS was used in bullying research, it was often modified and used on the local level so there are very few studies using the national version of the YRBS for bulling research.

Studying bullying among middle school children, using a modified YRBS, Pintado (2007) found that verbal bullying such as teasing and name calling, or physical threats was more prevalent than actual physical assaults or spreading rumors. Using the national 2003 YRBS data, Paxton (2007) found that engaging in risk behaviors increased the odds of depression. Although depression was reported more among females and Hispanics, controlling for demographics, adolescents who engaged in risk behaviors such as carrying weapons, physical fighting, substance use, and sexual intercourse had higher odds for depression (Paxton, 2007). In looking at the YRBS state level data, Peskin's (2007) found that bullied victims tend to internalize the problem (anxiety, depression) during middle and high-school years but bully-victims' internalization stopped after middle school. Race and gender also lead to more of a feeling of victimization. According to Fitzpatrick (2010), using variables similar to YRBS, they found

African-American youths had higher levels of depression, no matter the label (e.g., bully, victim, or bully-victim), and had a higher risk of exposure to violence.

Teen Alcohol abuse and use

The American Psychological Association's (APA) manual of mental health disorders, Diagnostic and Statistical Manual of Mental Disorders-IV (DSM-IV), has two definitions used to describe people who have Alcohol Use Disorder (AUD): Alcohol Abuse and Alcohol Dependence. You are categorized as an alcohol abuser if one of four criteria is met:1) alcohol interferes with fulfilling life responsibilities; 2) alcohol use causes engagement in hazardous behavior such as drinking and driving; 3) alcohol use has caused legal issues; 4) use of alcohol has contributed to social and relationship problems. In contrast to alcohol abuse, alcohol dependence has more of an internal component that causes physiological symptoms. Someone is alcohol dependent if they have built a tolerance to alcohol. To be tolerant to alcohol means in order to achieve the same level of intoxication one must increase the amount of alcohol consumption; the existence of withdrawal symptoms, such as shaking, that usually take place four to twelve hours after the absence of alcohol; the experience of physical and psychological problems; frequent unsuccessful attempts to quit; planning life around alcohol usage; or drinks longer and in greater quantities (APA, 2000; Chung, 2006).

Teen alcohol usage. There are problems with the definition of the alcohol use disorders when applied to adolescents. Winters (1999) found that many adolescents were more often classified as alcohol abusers than dependents. Ellickson et al. (1996) found that only less than 4% of the high school seniors who wanted to stop drinking could not, whereas, 24% were more likely to have passed out because of drinking. An explanation of the problem of generalizing this definition for every population can be explained by the sample used to define alcohol use disorder. Cottler et al. (1995) found that only one of the 17 DSM-IV field trials contained adolescents (as cited in Deas, 2000). Because of this problem, DSM-IV does not describe general alcohol involvement for adolescents (Chung, 2006). Researching the difference between alcohol use between adults and adolescents, Deas et al., (2000) found that adolescents drank less frequently than adults, were less prone to blackouts, and that there were no differences in the quantity of alcohol consumed per occasion. In general, teenagers are still developing into adulthood, so the physiological and psychological differences that exist are compounded when alcohol is interjected. So, the definition of alcohol use needs to be altered to account for teen behavior.

During adolescents, children are still developing. Adolescents experience a change in sex hormones, estrogen, testosterone, and an increase in growth hormone, as well as, the stresses of novel life stressors, such as adult/life choices (USDHHS, 2006). Consuming alcohol during this period upsets the hormonal balance needed for normal growth development for organs, muscles, and bones (USDHHS, 2006). Excessive drinking by teens causes problems with short-term and long-term memory. Research has found that teens consuming over 21(males) or 14 (females) units of alcohol per week had flaws in long-term and short-term memories (Heffernan, 2005). A unit was described as a half of pint of beer or a small glass of wine (Heffernan, 2005). The reason for the difference is that alcohol is water soluble, since females are have more fat and less water, females tend to feel the effects more readily (IBC Cutting edge, 2002; Lopez & Kelly, 2002). In general, the effects of alcohol depend on several variables including the gender, age, emptiness of the stomach, weight, and concentration of alcohol consumed (Lopez, 2002). Overweight teens, even with moderate use of alcohol, showed an increase of liver enzyme that was an indication of liver damage (USDHHS, 2006). Impaired cognitive abilities, such as, problem solving and concentration, is another known effect of alcohol use in teens (Biddulp, 2005; Evashwick et al.,1998). Also, adolescents who drank frequently or had multiple drinks per occasion also had an increased dropout rate between 9 to 11 percent (Chatterji & DeSimone, 2005).

Using the definition of an alcohol user by the United States Department of Human and Health Services (USDHHS) (i.e., an alcohol user is someone who drinks at least one drink per month) Maney (2002) studied the interaction of risk behaviors and teen alcohol users. Defining a high-risk alcohol user as someone who binge drinks once or twice a year, low-risk alcohol users as someone who does not binge drink, and binge drinking as four or more drinks at one occasion in the past 30 days, Maney (2002) found that high-risk drinkers were more likely to be report regret after drinking; combine drinking with sexual encounters; fighting; and report problems with friends, family, and in school (Maney et al., 2002). Maney (2002) used the USDHHS's 1996 National Household Survey on Drug and Health (NSDUH) to define binge drinkers and alcohol users. According to the NSDUH (2009), there are three categories of drinkers: (a) current users are those who had at least one drink in the past thirty days; (b) binge drinker are those who had five or more on the same occasion at least one day within the past thirty days; and (c) heavy users as those who have had five or more drinks on the same occasion on each of the five or more days in the past thirty days. According to their statistics, the 2009 survey, only 15% of youths between age 12 and 17 were current drinkers, 9% binge and 2% heavy drinkers. Whites were more likely to currently use alcohol than any other group; they were followed by Blacks, Hispanics, Asians, and lastly, Native Americans. Compared to YRBS, NSDUH surveys a larger populous, twelve years or older versus 9th to 12th graders. The YRBS (2009) showed that 41.8%

of students had at least one drink during the past 30 days, 72.5% of all students have had at least one drink of alcohol in their life time, and 24.2% have had five or more drinks in a row on one day during the 30 days before the survey. The slight differences in the numbers could be because the age range and the fact that, according to Muthén and Muthén (2000), alcohol use tends to increase with age (as cited in Mason et al., 2010). Age of first use predicted progression to alcohol related disorders with adolescents between 11-14 years old showing an increased risk (DeWitt et al., 2000). Nevertheless, both surveys show adolescents are consuming alcohol, and because there are health and social effects of alcohol use, it is imperative to know how to describe and better define alcohol use during adolescent.

Ellickson and Hays (1991) described problems with alcohol in middle schoolers, and seventh and eighth graders in terms of frequency of drinking and heavy drinking. Heavy or binge drinking was defined by the number of days the student consumed three or more drinks at one setting. Expounding, Hays and Ellickson's (1996) research defined alcohol misuse in teens as the following three dimensional construct: 1) frequency and quantity as one dimension; 2) high-risk drinking; 3) negative consequences. Frequency was defined as number of drinks in the past month or year. Quantity was the average number of drinks per day. High-risk drinking measures how alcohol affects behavior or judgment such as drinking and driving, binge drinking, public intoxication, intoxication at school, and use of other drugs. Negative consequences consist of internal and external consequences for those who use alcohol such as feeling sick, getting into fights, missing school, being arrested, and having accidents while driving (Hays & Ellickson, 2006). Reboussin and colleagues (2006) researching the types of underage drinkers, found that there were three classes of underage drinkers: non-problem related drinkers (i.e., drinkers with no alcohol-related problems); risky problem drinkers (i.e., drinkers who experience physical problems from drinking or have alcohol-related social consequences); and regular problem drinkers who have greater severity of alcohol-related social consequences (Reboussin et al, 2006). Reboussin et al. (2006) used latent class analysis, which categorized based on percentage of severity, to identify students with alcohol problems while Ellickson and Hays (1996) used cut-off scores based on the grade-level to identify teens who misused alcohol.

According to Nansel et al. (2004), bullies, victims, and bully-victims all had emotional, social, and health problems, but bully and bully-victims were more likely to use alcohol and to carry weapons. Like other research, victims had more emotional and social difficulties and bullies reported more school difficulties. Bully-victims were a cross between the two in terms of outcomes. Bully-victims had similar emotional and relationship issues as victims, whereas they shared school difficulties and alcohol use issues with bullies. Health issues were relatively equal in all groups.

Suicide literature review

How can killing oneself be complex? Scenario A: Susie had a habit of cutting herself. One day, she cut an artery and died. Her family said she was known to suffer from depression from time to time. Scenario B: Robert loved to work hard and play even harder. While copying his friends' jumping moves, Robert tried to outperform his friends by jumping a large hill; he fell head first on the cement and died. Scenario C: John used to tell people he wanted to die but he had recently received psychotherapy which helped him recover. All three scenarios could be attributed to suicide or suicide attempt. The answer to which one of these people committed suicide is challenging to answer.

Prior to the mid 1980's, there was no formal agreed upon definition of suicide. Without any kind of a rubric, coroners and medical examiners had to determine the cause of death (O'Carroll, 1996). During the mid-1980's, the Center for Disease Control (CDC) assembled researchers, medical professionals, and statisticians to define suicide. According to Rosenberg et al. (1988), the outcome of their collaboration was a definition of suicide, the Operational Criteria for the Determination of Suicide (OCDS), which stated suicide was "death from injury, poisoning, or suffocation where there is evidence (either explicit or implicit) that the injury was self-inflicted and that the deceased intended to kill himself or herself" (as cited in O'Carroll, 1996, p. 246). Despite this, there was still no consensus on the definition of suicide.

Ivanoff (1999) defined suicide as self-initiated and intentional death. Mayo (1992) definition stated that suicide had four components: "[S]uicide has taken place if death occurs, it must be of one's own doing, agency of suicide can be active or passive, and it implies intentionally ending one's own life" (p. 92). Silverman and Marris' (1995) definition on suicide stated, "Suicide is, by definition, not a disease, but a death that is caused by self-inflicted intentional action or behavior" (p. 522). Lastly, the World Health Organization (WHO) (1998) defined suicide as "[t]he act of killing oneself deliberately initiated and performed by the person concerned in the full knowledge or expectation of its fatal outcome" (as cited in De Leo, 2006, p.8).

These definitions define suicide not suicide attempts. O'Carroll (1996) argued that Rosenberg did not operationalize suicide but merely gave a definition. He wanted to operationalize the term suicide by proposing a nomenclature; he defines a nomenclature for suicide as a clear, unambiguous, and basic terms associated with suicide (O'Carroll, 1996). Using a topology, he defined a suicidal act as "potentially self-injurious behavior for which there is evidence that the person intended to kill himself/herself" (O'Carroll, 1996, p. 247). O'Carroll (1996) included cases where death did not occur, considered evidence of intent which resulted in a more comprehensive definition of the term suicide. However, some people took issue with the word intent being in the definition of suicide. How can a dead person tell you they intended to kill themselves? Was John intending to kill himself? If someone survives the suicide attempt, do they themselves know their own intent? De Leo (2006) noted that some people have no intention of harming themselves but seek attention; intention is a construct has not been defined and only inferences can be made on intent, and even if the person survives the suicide attempt, you are assuming the person is aware of their actions and will have memory of what occurred. Based on WHO (1998) definition of suicide, De Leo (2006) proposed a definition of suicidal behavior that describes suicide as "non-fatal" and "fatal" behavior with or without injuries.

Silverman et al (2007), revisited O'Carroll's (1996) definition of suicidal related behavior by renaming some categories and adding an additional category, undetermined. According to De Leo (2006), the problem with most definitions of suicide is that some definitions have cultural judgments. Does cutting oneself mean you're suicidal in Africa, especially when marking or cutting is a cultural tradition? De Leo (2006) noted that there needs to be universal definition of suicide. To address these issues, the WHO said suicide should be defined as an act with fatal outcome for which the deceased knowing or expecting a fatal outcome initiated and carried out for that purpose (De Leo, 2006).

Studying commonalities of those who purposely commit suicide, Shneidman (1996) found that those who commit suicide see suicide as a solution, use suicide as a way to end the consciousness of pain, have an unmet psychological need or want, express feelings of hopelessness, are usually cognitively ambivalent about the value of living, commonly perceive lack of options, see suicide as an escape, usually leave clues or communication of their intentions or distress, and have a tendency to have a limited view on how to cope in life (Shneidman, 1996). Hosansky (2004) describes some common warning signs for those who are about to commit suicide as: committing self-harm; obsessing about death; writing about death; change in personality, behavior, eating or sleeping patterns; feelings of guilt; and decreased academic or work performances (as cited in Rudd et al., 2006). According to Rudd et al. (2006), warning signs have an underlining proximity attached that suggest the risk of suicide is immediate. Conversely, simply looking at factors that contribute to suicide such as, history of mental illness, past history of suicide, and stressful life events are only long range assessors of risk (Rudd et al, 2006). Baldessarini (1988) noted that in general it is difficult to predict who will commit suicide because of the low rate of suicides, but despite the false positives, it better to error on the side of caution (as cited in Rudd et al., 2006). For teens, the most common factors were previous suicide attempts, using substances, worrying about depression, and failing in school (Hacker et al., 2006). Hendin (2001) found only a small percentage, of the small number of suicide cases, had actual intent to kill themselves; so focus should be on immediate signs of risk, which he identified as someone having a life event; depression mixed with a current state of desperation, guilt, or rage; and a observable risk behavior (e.g., talking about suicide, social deterioration, increase substance abuse).

King and Apter (2006) noted that surveys designed to measure the prevalence of suicidal ideation (suicidal thoughts) and behaviors are notably hard to interpret because they generally ask only if they have attempted suicide, but there is a lack of consensus on the definition of suicide attempt because some believe it implies intent. To judge severity of the attempt, King and Apter (2006) believes the question should be rephrased as "have you tried to take your own life, which implies intent". In general, it is a good rule of thumb to apply more than one definition when asking about suicide attempt, and to specify the "outcome of the attempted suicide through a serious of increasing specific questions"(p.64), such as asking first about injury, then hospitalization, and lastly, medical attention (King & Apter, 2006). YRBS is one of the few surveys to associate medical outcomes to suicide attempts (King & Apter, 2006).

Alcohol use is a risk factor for teens (Hacker, 2006). Schilling et al. (2009) found that alcohol use was a key factor between those who planned and those who did not plan to commit suicide. Planned suicide attempts were associated to high levels of hopelessness and depression, whereas, unplanned suicide attempts were associated with high levels of aggression, alcohol intoxication, increased aggressive and negative emotions, and decreased mental capacity to think of alternative coping strategies (Schilling et al., 2009). Unplanned suicide attempts were more prevalent in early adolescence and among males (Schilling et al., 2009). During stressful life events, adolescents are more likely to be impulsive (Schilling et al., 2009). The combination of stressful life events, age, and alcohol use exhibits two major warning signs. Hinduja (2010) found that bullied adolescents had an increased risk of suicidal thought. Ivarssoon (2005) found 47% of Bully-Victims, and 39% of victims reported having suicidal thoughts, compared to 12% of bullies. Only 12% of the bully-victims and 11% of the victims had seriously thought about suicide as a viable option (Ivarssoon, 2005).

Structural equation modeling (SEM)

Structural equation modeling is a multivariate statistical technique that encompasses ordinary linear regression, non-linear regression, and factor analysis (confirmatory). SEM has the ability to model observed, as well as, unobserved/latent variables. Observed variables are variables that are measureable such as weight. We can ask someone their weight or physically have him/her step on a scale. Happiness or Depression or Bullying is something that is difficult to define and is not directly measureable. With SEM we are able use theory to define those latent variables and perform statistical analysis on them, as well as, on their observed analog. When SEM is performed on observed variables it is called Path Analysis. Path Analysis suffers the same constraints as does Regression; it assumes variables have no measurement error. However, this assumption is a fallacy, because the assumption of no measurement error has rarely been found to be true (Duncan, 1975; as cited in Bedelian et al., 1997). When measurement error is present there is extra variability in the regression model not explained by independent variable(s). For example, the ordinary least square (OLS) regression modeling weight and hours of physical activity might seem like a simple model but it might not be that straightforward depending on how one measures the variable weight. One would expect a weight decrease when physical activity increases; so physical activity explains most of the decrease seen in weight. However, if there is indeed extra variability between weight and physical activity due to the measurement error then there is extra variability about the regression line. The unexplained variance that is not explained by weight goes into the calculation of the coefficient of determination as unexplained variance, hence, lowering the real impact that one expects weight to have on hours of physical activity. In essence it lowers the correlation between weight and physical activity, and attenuates or lowers the beta (parameter estimate; DeShon, 1998).

Spearman (1904) formula for true correlation is the Pearson correlation corrected by the square root of the reliability.

$$\rho_{xy_c} = \frac{\rho_{xy}}{\sqrt{\rho_{xx}}} \qquad (1)$$

Once we correct for reliability, the true correlation between the two measures are apparent. For example, if the uncorrected correlation between weight and physical activity is .50 and the reliability of physical activity is .70, then the corrected reliability .60. The presence of measurement error underestimates the reliability between the two measures, and since correlations are a function of the beta coefficient it attenuates the beta coefficient. In classical test theory (CTT) there are many ways to obtain the reliability such as test-retest, internal consistency, or alternate forms. The most popular one is coefficient alpha which can be conveniently done in one test setting and with one form (Bedeian et al., 1997). By including more than one variable as a definition of a construct, SEM has the ability to calculate reliability, and include it in the statistical analysis by correcting the Pearson correlation, and thereby parameter estimates.

Figure 1.1

SEM Model

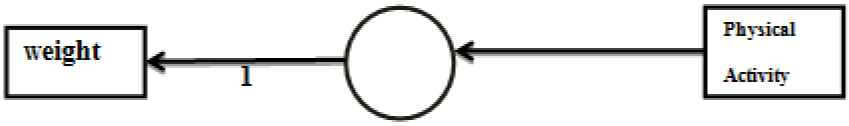
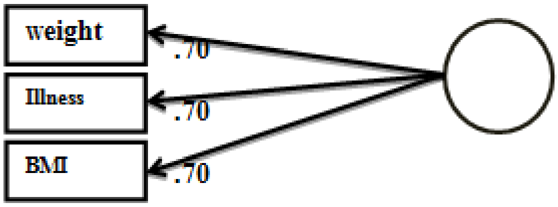


Figure 1.2

SEM Model 2



Instead of using one indicator such as weight, as in Figure 1.1, SEM incorporates several correlated variables that might cause problems with multi-colinearity in multiple linear regression Figure 1.2 shows an alternative model where these three indicators can be used for a fuller picture of general health. By making assumptions about the loadings, the error (tau or parallel equivalency), and unidimensionality, Cronbach alpha can be calculated from the confirmatory factor model presented. According to Deshon (1998), reliability for a composite variable was presented by Werts, Linn, and Jöreskog in 1974:

$$\rho_{xx} = \frac{\left(\sum_{i=1}^{p} l_i\right)^2}{\left(\sum_{i=1}^{p} l_i\right)^2 + \sum_{i}^{p} E_i} = \frac{(.70+.70+.70)^2}{(.70+.70+.70)^2 + (.30+.30+.30)} = .83 \tag{2}$$

Besides correcting for reliability, SEM is useful for mediation analysis. Mediation analysis can be done in regression or SEM. Aside from being able to incorporate latent variables, correcting for measurement error, and test competing models, mediation analysis in SEM is another big advantage of SEM because it allows one to run simultaneous equations in one single analysis.

Mediation Analysis

Mediation analysis is a causal model that seeks to explain why there is a relationship between the independent and dependent variables. If X causes M and M causes Y, then M is a mediator. A more formal definition was given by Barron & Kenny (1986), “a variable may be said to function as a mediator to the extent that it accounts for the relation between the predictor and the criterion” (as cited in MacKinnon, 2008, p.8). Mediators speak to how or why such effects between the independent and dependent variables occur just like a confounding variable. However, although a confounding variable might sound similar in definition, it is not the same as a mediator. A confounder explains the relationship between X and Y because it is related to both X and Y. A mediator explains the relationship between X and Y because it transmits the effect of X on Y (have to go through the mediator) (Mackinnon, 2008).

Figure 1.3

Confounding Model vs. Full Mediation Model

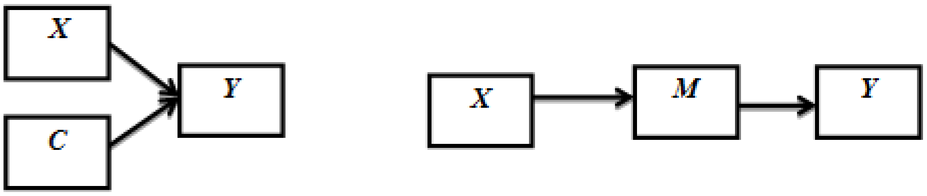


Figure 1.3 illustrates the path analysis of regular multiple regression and mediation analysis. In the confounding model the relationship between X and Y is explained by controlling for the confounding variable (C). In the fully mediated model, there is no direct relationship between X and Y, their relationship is a byproduct of their relationship with mediator (M). In the classic example of shoe size and intelligence, age mediates the relationship between these two variables. Note there may still be a relationship between shoe size and intelligence, so in this case age partially mediates the relationship between shoe size and intelligence (the correlation seen between shoe size and intelligence share a common cause). Partial mediation is found when there is a direct relationship between the independent and dependent variable as well as a mediator (figure 1.4).

Figure 1.4

Partial Mediation Analysis

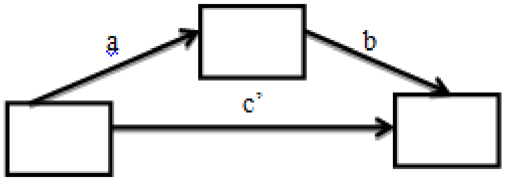
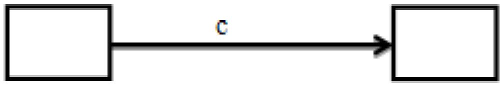


Figure 1.5

Direct Effect of X on Y



Baron and Kenny (1986) outlined a series of tests to perform in order to test for mediation. The three tests coincide with the regression equations formed when doing a mediation analysis.

$$M = i_1 + aX + e_1 \tag{3}$$

$$Y = i_2 + c'X + e_2 \tag{4}$$

$$Y = i_3 + cX + bM + e_3 \tag{5}$$

Each of these regression equations must be run separately since in regression no variable can serve as independent and dependent variable at the same time. In order to prove that X affects Y through a mediating variable M, according to Baron and Kenny (1986):

> A variable functions as a mediator when it meets the following conditions: (a) variations in levels of the independent variable X significantly account for variations in the presumed mediator (i.e., Path *a*), (b) variations in the mediator significantly account for variations in the dependent variable (i.e., Path *b*), and (c) when Paths *a* and *b* are controlled, a previously significant relation between the independent and dependent variables is no longer significant, with the strongest demonstration of mediation occurring when Path *c* is zero…..When Path *c* is reduced to zero, we have strong evidence for a single, dominant mediator. If Path *c* is not zero, this indicates the operation of multiple mediating factors (p.1176).

In practice, researchers first determine if a significant relationship between the independent variable and the dependent variable exists (Path *c*, which is called the total effect). Once that initial relationship is confirmed, the research then confirms if paths *a* and *b* are significant. Path *a* represents the parameter relating the mediator to the independent variable. Path *b* is part of the multiple regression equation so it represents the relationship between the mediator and the dependent variable controlling for the effect for the independent variable in the model. The product of *a* and *b*, *a**b,* represents the mediated effect or indirect effect. Path *c*' represents the partial affect for the mediator or the relationship between the independent and dependent variables controlling for the effects of the mediator in the model. This path is also referred to as the direct effect (Mackinnon, 2008). Establishing significance of the direct effect differentiates between a full or partial mediation. If the direct effect is not significantly different from zero then we have full mediation. The mediated effect, *a**b,* is also equal to the difference between *c* and *c*', *c-c*'. The total effect, *c,* is equal to the direct effect, *c*', plus the mediated effect, *a**b*= *c-c*'.

Baron and Kenny recommended an alternate single test to establish the significance of mediation effects, the Sobel (1982) z-test (Baron & Kenny, 1986).

$$z = \frac{a*b}{\sqrt{b^2 {s_a}^2 + a^2 {s_b}^2}} \tag{6}$$

One can look to see if the z value is signicant using the standard normal table or see if the confidence interval contains zero. Since a*b is equal to c-c', the Sobel test tests if the difference between the total effect and the direct effect is significant (Note, s_a^2 is the square standard error of a, s_b^2 is the square standard error of b). The assumptions are the same assumptions for regression, namely, linearity, no ommitted influences, no measurment error in (X, Y, nor M), normality, and the residual is uncorrelated with the predictor in each equation (MacKinnon,2008, p. 55). Note, logistic regression is applied when the independent variable is categorical.

Figure 1.6

Mediated Model Examplar

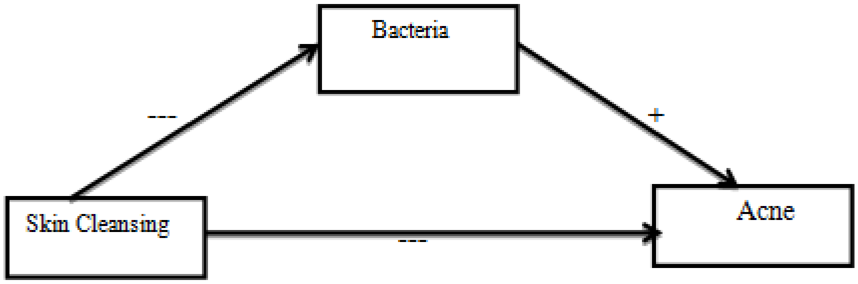


Figure 1.6 shows a typical example of a mediated model where we first establish that washing one's skin on regular bases reduces acne, but the reason for this reduction is the reduction of bacteria on the skin. The signs in the mediated model are similar to regression parameters signs. For one a unit increase in bacteria there is an increase in acne. For a one unit increase in skin washing or cleansing then there is a decrease in bacteria. Overall, the mediated effect, a*b or negative * positive, shows an overall negative effect on Acne through reduction of bacteria.

Recent Research. Recent research on mediation analysis has found that establishing an initial total effect, c, is not necessary. Rather, only a significant indirect effect is needed, a*b (Rucker, Preacher, Tormala, and Petty, 2011). Causing researchers to rely on this initial step has historically caused researchers to prematurely terminate viable research projects (Zhou, Lynch, and Chen, 2010). According to Rucker et al. (2011) there are many other outside influences that can explain the presence or absence of the initial direct effect, specifically, measurement precision, strength of relation, sample size, size of total effect, and suppressor variable.

When X and Y are moderately reliable, but M is highly reliable, then the power for any regression weight associated with a and b is increased. Therefore, due to this measurement imprecision one is more likely to detect an indirect effect, a*b, than a direct or total effect, c or c’. If the strength of the relationship between M and X is stronger than X and Y then a*b might have a stronger indirect effect even if c is not as strong. Mediation is dependent on sample size. As sample size increase you are more likely to final a total effect is present. In fact, given a small sample size you are more likely to label a mediation as a full than partial effect. The smaller the total effect, c, the more likely one is to detect a full mediation. Lastly, the absence of suppressor variables (i.e., variables that undermine the total effect, or the relationship between X and Y, by omitting it), when included in the model, is said to be controlled for and to strengthen the relationship between the independent variable X and dependent variable Y. The absence weakens the effect of X on Y. The omission of suppressor variables is another case when the total or direct effect can be non-significant. The hallmark of a possible suppressor variable is when the direct effect, c’, is the opposite sign of the mediated effect, a*b. This scenario is called inconsistent mediation.

Zhao et al. (2010) outlined new decision criteria for determining mediation and mediation type. The new decision rules looks solely at the mediation effect and the sign of the mediation. They concluded that there are four possibilities when doing mediation analysis: no effect, indirect only effect, competitive mediation, and complementary mediation. The total mediation equation is given by c= a*b +c’. If the mediation effect, a*b, is significant, then mediation is established but there are now different types that are possible. If the mediation effect, a*b, is not significant, then depending on if there is a total effect, c, of X on Y, then one either has missing mediators or problems with the theoretical framework (hypothesized mediator not identified) or both. If the mediation effect is significant then the sign of direct effect and indirect effects becomes important, as well as, the significance of the total effect.

If the total effect, c, is not significant but the mediation effect, a*b, is significant then we have an indirect-only mediation present. The hypothesized mediator is identified and there is low possibility of any other omitted mediators. If the total effect, c, is significant, then the sign of total effect equation signifies the type of mediation present, complementary or competitive. The significance of total mediation, c, specifies the presence of other possible mediators. Complementary mediation is when the mediated effect, a*b, and direct effect, c', both exist and are the same sign. Competitive mediation occurs when the mediated effect, a*b, and direct effect, c', exist and are opposite signs. The presence of these two types of mediators means an incomplete theoretical framework or the presence of other mediators in the model that has been omitted from the model (Zhao, Lynch, and Chen, 2010, p. 201).

In addition to changing the order of the significance testing, an alternative test to the Sobel z test was also suggested. Although rarely used in mediation analysis, the Sobel test requires a large sample size and an assumption of normality about the indirect effect, a*b, and sampling distribution (Preacher, and Hayes, 2004, p. 719). Preacher and Hayes (2004) bootstrapping test for mediation effect is a non-parametric test that is based on resampling with replacement of the variables involved in the mediation model. Unlike the Sobel test, it does not require normality of the sampling distribution of the indirect effect in the mediation model.

MacKinnon, Lockwood, and Williams (2004) found this bootstrap test for mediation affect to have more accurate confidence limits and better type I error rates than the Sobel test.

SEM versus Regression mediation analysis. The question about whether to do mediation analysis as a linear/non-linear regression model or SEM module might still remain. There are a number of reasons to do mediation analysis in SEM over doing it as a regression model. Namely, SEM can handle the issue of inaccurate measurement or measurement error. According to Hoyle and Kenny (1999) when measurement error is present in the mediator then this can lead to an attenuated relationship between the mediator, M, and the independent variable, Y (as cited in MacKinnon, 2008). The most notable benefit of using SEM over regression for mediation analysis is the reduced standard error. When running a standard mediation analysis (i.e,. mediation analysis with one mediator), you have a minimum of three equations that have to be run as three separate regression equations. SEM runs all equations simultaneously which reduces standard error because all the parameters are in the model. Missing parameters by fitting individual equations means more unexplained influences in each model are let out (Iacobucci, 2008, p. 22).

CHAPTER THREE

STUDY ONE METHODS

Youth Risk Behavior Survey (YRBS), the survey most commonly administered to American teens, employs a complex sampling design which is both time and cost efficient for the Center for Disease control (CDC) who created the survey. This chapter describes the design and research methodology that was implemented to study the relationship between bullying and suicide using data obtained from the 2009 YRBS. Described in this section is the YRBS data set, variables used in the study, descriptive statistics on the variables involved in the study, the statistical model, rationale for the model, and software used to run the model. There are three basic research questions I want to answer using the mediated multilevel SEM framework:

1) Is there a relationship between bullying and depression/suicidality?

2) Is that relationship mediated by school violence?

3) Is teen alcohol abuse and use another possible mediator in the model?

YRBS 2009 Data

Overview. YRBS is the most common survey used to collect health related data for teens in schools. The survey's main focus has been questions related to violence, suicide, substance abuse, weight issues, and sexual behavior. The 2009 national version of YRBS was the first time that a question on bullying was included on the survey. The recent 2011 YRBS included a second question on bullying, namely, online bullying. Because of this, the few researchers who have studied bullying using YRBS had to modify the YRBS and reissue a smaller state level version of the survey (Pintado, 2007). According to my research, no one has used the YRBS national data set in relation to bullying and suicide nor has anyone thought to use SEM. This could be because CDC YRBS literature advises researchers how to perform logistic regression on YRBS. Indeed, most of the research done on YRBS has been done using logistic regression. Kupek (2006) noted that it is rare for SEM to be used on medical related data, logistic regression is quite common. However, bullying is a construct that is not perfectly measured by one question. This unnecessary inclusion of a variable that is assumed to be perfectly reliable attenuates the beta relationship, if found, and also adds more variability in the unexplained variation between bullying and suicide. Although I am faced with this same limitation without information on the reliability of bullying, I can study the other variables as constructs which helps reduce some of the unexplained variance.

Subjects and instrument. The 2009 YRBS questionnaire sample consisted of 16,410 students, grades 9 through 12, from 158 participating public and private schools. Using the software program, Mplus 6.1, I found there were 16,409 subjects and 55 clusters instead of the 57 primary sampling units or counties given in the YRBS documentation. The average cluster size was 298.345 (55 times 298.345 gives you 16,408.975 or 16,409 subjects). The PSU consisted of counties or groups of counties not schools.

The national version of YRBS used a three stage complex sampling design. During first stage of the design, about 1200 large-sized counties or groups of “smaller adjacent” counties are selected (CDC, 2009, p.2). From these, counties are subdivided into 16 non-overlapping groups or strata based on their density and ratio of minorities. Roughly 57 counties are then selected with school enrollment size as the probability of selection. During the second stage, almost 200 public and private schools having grades 9th-12th, are selected from the 57 counties based on their school enrollment size. Schools with a higher enrollment of minorities are sampled at a higher rate. The third stage consists of a random sample of one or two required 9th-12th classes within the selected school. The entire class is selected to be in the study, students have the option to not participate.

To adjust for oversampling minorities, non-responses, and the unequal probability of selection, YRBS employs a weighting of each student to keep the sample more representative of the target population (i.e., 9th to 12th grade students). After which, an overall weight is then applied so that the total number of students in the sample adds to the sample size.

From year to year, the YRBS survey does vary slightly in the number of questions being asked and content, depending on the national objectives at the time. Most questions are ordinal in scale; however, there are a small number of dichotomous variables. Most of the ordinal variables have at least five progressive categories. Along with the health-risk questions, demographic background information is collected such as gender, age, grade level, race, and geographic region.

Variables. Although the 2009 YRBS data set consisted of 87 standard questions, of which 11 questions were added, only nineteen variables related to my research (CDC, 2009). Of the 19 variables, 16 observed variables were used to conduct the analysis. The additional variables added in the data files (i.e., weight, stratum, and psu) were not included in this count. As appendix Table A.1 indicates, most of the observed variables are ordered categorical with a few being dichotomous. Each ordered categorical variable has at least five categories, but most, if not all, are on different scales.

Alcohol abuse in teens can be described by increased risky behavior, frequency of use, and quantity of use. The 2009 YRBS data set had six direct questions concerning alcohol and teens that the CDC labeled as alcohol usage. I selected five of the six questions from the usage category, ignoring the questions about the age one began drinking and where one purchases alcohol. I then added another question from the safety category that was related to alcohol and safety as well as drinking and driving. I labeled this unobserved variable as *teen alcohol abuse and use* that has a total of five indicators: 1) number of times student drove drunk in the past 30 days, 2) average number of days student had at least one drink of alcohol, 3) during the past 30days, number of days student had at least one alcoholic drink, 4) during the past thirty days, number of days student drank five or more drinks on one occasion, 5) in the past 30 days, number of days the student had at least one drink on school property. Drinking at school and drinking while driving are indicators of risk behavior, average number of days for drinking at least one glass of alcohol measures the frequency, and the number of days having at least thirty glasses of alcohol measures the quantity. Most of the questions, except the driving drunk question, are on a seven point ordinal scale but not the same seven point scale. For example, question 42, number of days you drank five plus glasses of alcohol at one occasion, allowed the choice of zero, one, two, three to five, six to nine, 10 to 19, and 20 or more days; while question 41's, number of days had at least one glass of alcohol, choices were zero, one to two, three to five, six to nine, 10 to 19, 20 to 29, and all 30 days.

Bullying involves repetition, intent to harm, and exhibits an imbalance of power. Bullying can be either direct or indirect. It takes the form of physical aggression, verbal abuse, and social/relational harm (willful intent to harm peer relationship such as gossip). YRBS has no school level variables, and only one question asking directly about bullying. YRBS had ten questions labeled violence related questions. Violence is one dimension of bullying that takes the form of physical aggression. Because YRBS did not have the full dimensionality of bullying, I decided not to create a latent variable called bullying; instead I used school violence as a latent variable to explain the relationship between bullying and suicidality. Of the ten possible questions on violence, only four related to school and violence: number of days you brought a weapon on school property, number of days felt threatened or injured with a weapon on school property, number days missed school because you felt unsafe, and the number of times you were in a physical fight on school property. All the questions have at least five scale points and the one question of bullying is a dichotomous yes or no question.

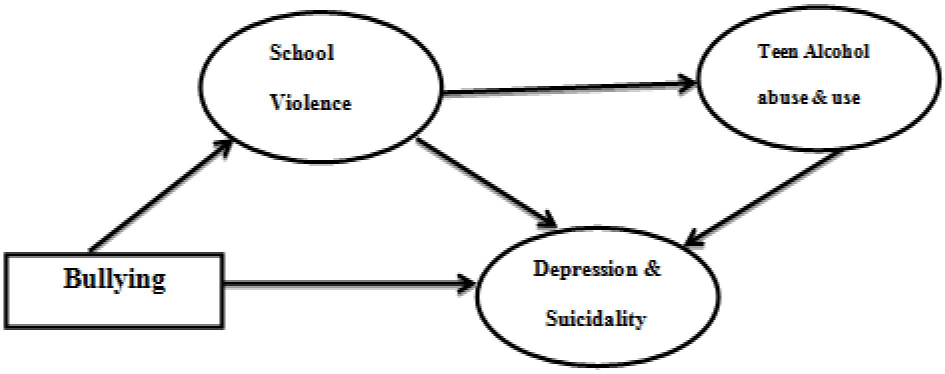
Teen suicidality is thoughts, plans, and attempts of suicide (Klomek, A., Marrocco, F., Kleinman, M., Schonfeld, I., & Gould, M. S., 2007). The 2009 YRBS asked five questions that they categorized as *feeling sad* and *attempted suicide*. Leaving out the question on medical attention after suicide attempt, four questions were used to define a latent variable depression/suicidality. These four questions were: 1) have you felt hopeless or sad for more than two weeks in a row within this past year? 2) have you considered suicide? 3) have you planned suicide? 4) and have you attempted suicide? Most of these questions were dichotomous.

Distributionally most of the data are positively skewed. Skew, mean, and standard deviation are all descriptors of ratio or interval data. With categorical data one usually uses median or percentiles to describe the data. Using Lisrel 8.80, I was able to obtain data frequencies. Looking at the percentages on categorical data, you see that most of the ordinal data are positively skewed (see Table A.2). Also, it should be noted for the polytomous responses the response categories went from negative to positive, meaning no alcohol abuse or no violence to increase frequency. The dichotomous questions, bullying and most of the suicide questions, were dichotomized as yes or no. These dichotomous variables were negatively skewed, assuming continuous data, with a high preponderance answering no.

Further analyzing the variables, we can look at the correlations between the variables. Instead of using a Pearson correlation matrix because our variables are categorical or polytomous indicators, the computer software running the analysis, Mplus, uses a polychoric correlation matrix (Brown, T. , 2006). The correlations were between .060 to .926, with the lowest correlations being between questions on teen alcohol use and either abuse and bullying at school or depression/sucidiality. All other correlations between the different variable types were low to moderate, see Appendix Table A.3.

Figure 1.7

The Mediated Model



The Model

Figure 1.7 outlines the mediated model that will be ran. *School violence, Depression,* and *Suicidality*, and *Teen Alcohol Abuse and Use* are all latent variables. The one dichotomous question on bullying will serve as an observed predictor variable that is mediated by *school violence*. Another competing mediator is included in the model, *Teen Alcohol Abuse and Usage*, to possibly explain the relationship between *School Violence* and *Depression/Suicidality*. Based on the rationale for testing the model (i.e., bullied adolescents have an increased risk of suicidal thoughts and adolescents spend most of their time in school), school violence may explain some of the reason we see this connection. Also because alcohol abuse and use is one of the key factors in teen suicide it was included as a possible mediator as well (Schilling et al., 2009). When we say a variable M (school violence) mediates the relationship between X (bullying) and Y (depression/sucidality), one goes from a simple correlation between X and Y to a partial correlation between X and, Y controlling for the effects of M. Due to controlling for the mediator or explaining some of correlations between X and Y by holding M constant, the relationship between X and Y will be reduced.

Software. Mplus 7 will be used to analyze the mediated SEM model. Mplus is a statistical software program specifically designed to analyze causal models involving observed and latent variables. The YRBS data are nested or multilevel by design. Because students within a school or class behave similarly this violates the requirement of independence. Non-independence is brought about due to the non-equal probability of selection. By design YRBS was created not as a random sample of students across America but as a complex three stage sample design to save time and money--this convenience for the designers leads to complexities for researchers. A Hierarchical or Multilevel model was designed to address these complexities. Mplus allows researchers to conduct multilevel modeling in two ways: modeling the between level structure and correcting for the standard error, or just correcting for the standard errors. Standard errors are underestimated in nested data because students in groups such as classrooms tend to behave similarly. Analysis on this type of data makes it looks like the scores are have less variance than is truly there. One of the benefits of doing a multilevel SEM model is that the between correlational structure can be modeled. However, there has to be enough variance to model the between level structure. When not enough variance between the clusters is present, Mplus allows one to run the model as a single level and corrects for underestimation of the errors by using the key word COMPLEX in Mplus code. The interclass correlation (ICC) is used to described the between level variance. Depending on the value of the ICC, modeling the between level structure is deemed possible or not possible. The ICC for the variables chosen in this analysis varied ranged from .008 to .04.   This indicated that there was not enough variance on the second level to consider modeling the second level. Therefore a standard SEM will be performed with the correction for the standard errors. When dependent variables are not continuous, they are defined in Mplus with the key words CATEGORICAL ARE, and the weighted least square with mean and variance (WLSMV) estimation method is used by default when running the analysis.

Missing data. There are three types of missingness referred to in literature: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Missing completely at random is missingness that has no rhyme or reason. Missing at random occurs when missingness can be explained by other variables, and thus missing values may be filled in via multiple imputation.. Missing not at random refers to the situation when neither MCAR nor MAR occur, meaning multiple imputation would not provide correct estimates of missing values. For example, a study about IQ and job performance is said to have missing at random pattern when only the low IQ workers have job performance scores missing, whereas, if workers with low job performance IQ scores are the ones with scores missing then we call this missing not at random (Enders, 2011).

There are four approaches to dealing with missing data: available case method (includes listwise and pairwise), single imputation method, model based imputation, and full informal maximum likelihood estimation (FIML; Kline, 2011). The available case approaches are where you delete incomplete cases. With listwise you delete cases where there is missingness in all the variables or complete missingness. Pairwise is less restrictive in that it looks at the covariance matrix and missing bivariate pairs are discarded (Enders & Bandalos, 2001). Single imputation replaces one missing score with an estimated score usually based on regression or group mean substitution. Model-based imputation method uses the model to generate more than one viable replacement for the missing scores (Kline, 2011). The FIML estimation approach does not delete data nor impute the missing data. Similar to a regression imputation, it uses an estimation method and also estimates the parameters and standard errors in one step. FIML uses the raw data (missing and all), means, and covariances to estimate these parameter estimates by maximizing a casewise function that incorporates all the information gathered. FIML estimation was specifically written for SEM applications (Graham, Olchowski, & Gilreath, 2007).

Each approach has their positive and negative attributes. Listwise and Pairwise deletion require MCAR assumption and yield biased parameter estimates under MAR, FIML yields more efficient estimates than listwise and pairwise under MCAR, and FIML parameter estimates are unbiased under both MCAR and MAR (Enders & Bandalos, 2001). Direct ML assumes data are MCAR or MAR and multivariate normal. It is assumed that FIML estimates can be biased MNAR. When data are non-normal, the robust maximum likelihood (MLR) estimator is used instead (Brown, 2006). According to the Mplus user guide, Mplus provides maximum likelihood estimation under MCAR, MAR, and MNAR for continuous and noncontinuous data (Muthén, & Muthén,2008- 2010, P. 7). For categorical outcomes, missingness is allowed for covariates but not outcomes when the weight least square (WLS) based estimator method is used. WLS based estimators are generally used when the number of categories are less than five. When WLS based estimators are used, a Bayesian analysis or multiple imputation method serves as an alternative to FIML (Muthén, 2008-2010). Bayesian analysis modeling of missing data gives asymptotically the same results as ML estimation under MAR (Muthén, 2008-2010, p.338). However, where there are large data sets this leads to large models with a large number of parameters which add a level of complexity that leads to non-convergence of the model (Asparouhou, & Muthén, 2010).

For my data set, the number of missing data range from 39 to almost 2000 with the questions related to attempted suicide having almost 2000 missing responses. According to initial analysis, using Lisrel 8.8 doing listwise deletion would drop the sample size down to roughly 12000. However, when you tell Mplus to correct for missing data by typing Missing in the analysis line and use WLSMV estimator, you no longer are using direct ML but pairwise deletion. If you have missing data and don't specify missing in the analysis line, listwise deletion is used. Pairwise deletion deletes fewer cases than the listwise. In fact, using pairwise in Mplus, the sample size used was at least 15633 depending on the model. When data are missing completely at random, pairwise deletion is better (Brown, 2006). For samples with a large N, pairwise deletion can lead to negatively biased standard errors, and if the data are MAR, the parameter estimates and standard errors are severely biased (Brown, 2006). MLR, robust ML, will treat the categorical variable as continuous, which is fine if the number of scale points is over five and the distribution is roughly normal, which does not apply here. There were two questions that have a high number of missing cases: 1) in the past twelve months how many times you have attempted suicide, 1801 missing 2) how many days has the student had at least one glass of alcohol in the past thirty days, 1546 missing. The rest of the missing is in the thousands. This could be a measure of response bias due to the nature of the question, According to Scafer and Graham (2002), in general, there is no way to test for MAR holds in a data set without follow-up data from the non-respondents or by imposing an unverifiable model. Because of the limitation of my data I am using WLSMV with pairwise elimination.

Limitations. The first notable limitation is the dearth of questions on bullying. Asking someone if they have or have not been bullied on school property does not capture the totality of the measure. Everyone has experienced some type of bullying at one time or another but without questions that measure frequency or intensity it is not known if the question is really measuring bullying. This relies on the person's memory and definition of bullying. No matter what analysis one chooses to do this would be a limitation. Second, I could not test to see if the same model holds for the between level because of the lack of variance or low ICC at the between level. Lastly, due the complexity of using data imputation on a large data set, pairwise deletion had to be implemented on data that could possibly be MNAR.

CHAPTER FOUR

STUDY ONE RESULTS

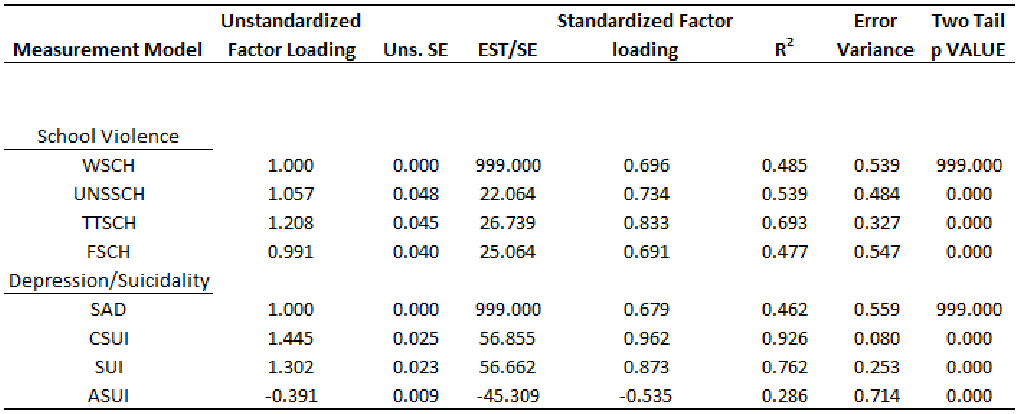
This section presents the results of the study about the mediated relationship between bullying and depression/suicidality using the Youth Risk Behavior Survey (YRBS). There are three questions that this section will answer: Is there a relationship between bullying and depression/suicidality? Is that relationship mediated by school violence? Is teen alcohol abuse and use another possible mediator in the mode? Each question is presented with the model and the interpretation of Mplus results. Following the results section, a discussion section delves deeper into the meaning of the outcomes and how it relates to previous empirical studies.

The first model ran was the core mediation model, figure 8, which tests if the relationship between bullying and depression/suicidality is mediated by school violence. The first step was to test the fit of the model. The most reported fit index is Chi-square which assess overall model fit. However, Chi-square is sensitive to sample size and will always report good model fit when the sample size is large; also, it assumes multivariate normality. The second most popular fit index is the root mean square error of approximation (RMSEA). RMSEA is a fit index that measure how close the implied matrix (model matrix) is from the observed variance covariance model. RMSEA is beneficial because it takes into account the complexity of the model by adjusting for the number of parameters. For RMSEA a zero means the model has perfect fit—with good fit being less than or equal to .05 (Hu and Bentler, 1999). There are numerous fit indices to choose from in SEM but each of them has their own positives and negatives. In general, when the WLSMV estimator is selected the comparative fit index (CFI), RMSEA, Tucker-Lewis non-normed fit index (TLI), and weighted root mean square (WRMR) should be used (Bowen & Guo, 2012). TLI and CFI recommend values of .95 or higher. WRMR requires values less than .90 for normal data and less than1 for non-normal data (Yu & Muthén, 2002; Bowen, & Guo, 2012,). The RMSEA=.023 (90% CI: .020, .026), CFI= .983, TLI=.976, and WRMR =1.625. Although WRMR showed the model was not an adequate fit, the other indexes show model fit well. The modification indexes recommendations did not agree with theory.

The factor loading represents the strength of the relationship between the variable/indicator and the common factor. The square of the standardized factor loading represents the percentage of variance in the indicator explained by the common factor. The error variance is the unexplained variance in the indicator not explained by the factor. For this model, the standardized factor loadings were all significant and ranged from -.533 to a high of .962.

Table 1.1

Measurement Model



Measurement Model	Unstandardized Factor Loading	Uns. SE	EST/SE	Standardized Factor loading	R^2	Error Variance	Two Tail p VALUE
School Violence							
WSCH	1.000	0.000	999.000	0.696	0.485	0.539	999.000
UNSSCH	1.057	0.048	22.064	0.734	0.539	0.484	0.000
TTSCH	1.208	0.045	26.739	0.833	0.693	0.327	0.000
FSCH	0.991	0.040	25.064	0.691	0.477	0.547	0.000
Depression/Suicidality							
SAD	1.000	0.000	999.000	0.679	0.462	0.559	999.000
CSUI	1.445	0.025	56.855	0.962	0.926	0.080	0.000
SUI	1.302	0.023	56.662	0.873	0.762	0.253	0.000
ASUI	-0.391	0.009	-45.309	-0.535	0.286	0.714	0.000

According to Table 1.1, there seems to be evidence that the construct *depression and suicidality* explains very little of the variance in the attempted suicide variable or indicator (ASUI) and should possibly be dropped from the model. Although not referenced in the modification index, the fact that 71.40% of the variance is left unexplained by the construct and that a number of missing responses makes response bias seem likely, dropping ASUI might be a good choice.

Figure 1.8

Core Mediated Model (standardized)

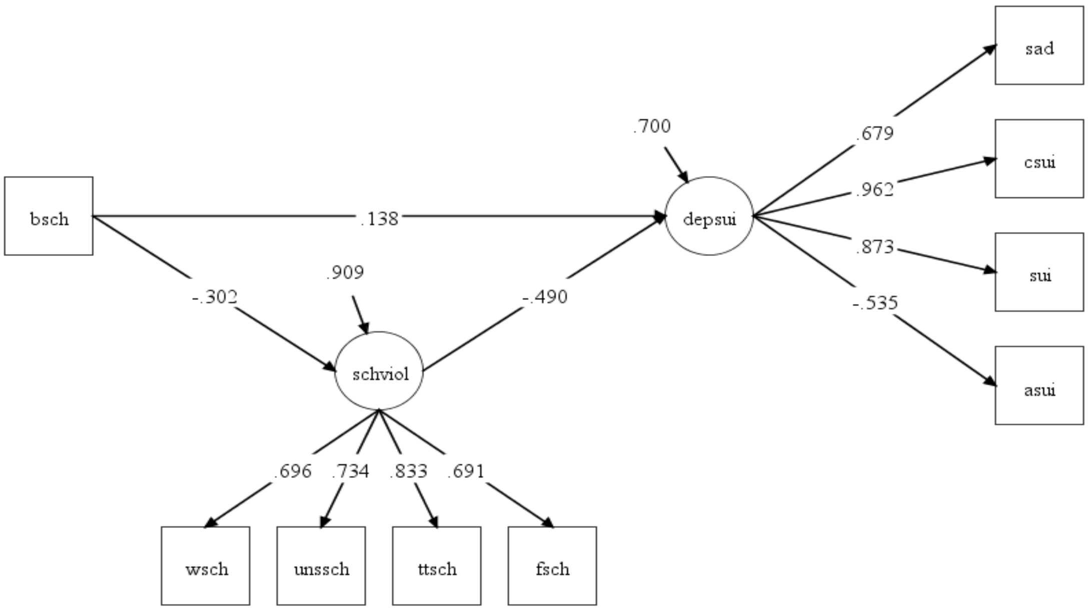
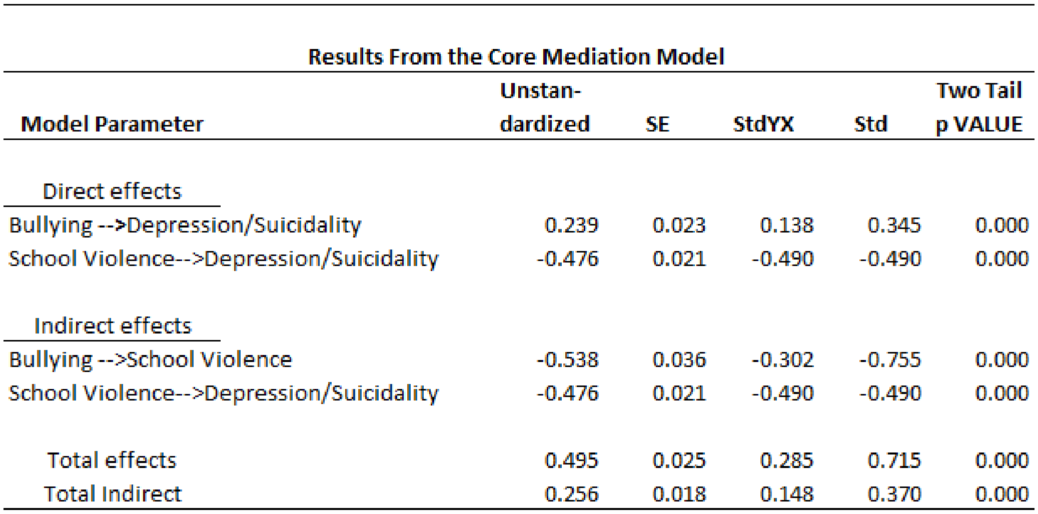


Table 1.2

Direct/Indirect Effects



Results From the Core Mediation Model

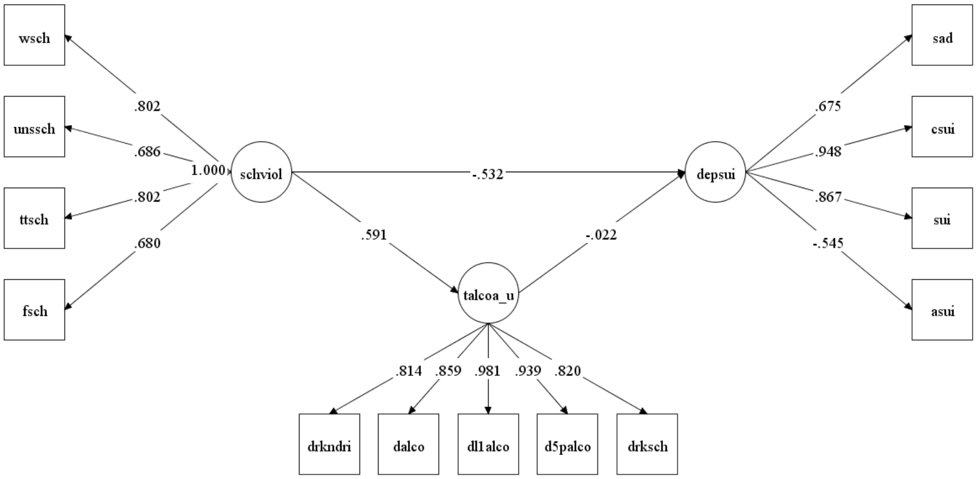
Model Parameter	Unstan- dardized	SE	StdYX	Std	Two Tail p VALUE
Direct effects					
Bullying -->Depression/Suicidality	0.239	0.023	0.138	0.345	0.000
School Violence-->Depression/Suicidality	-0.476	0.021	-0.490	-0.490	0.000
Indirect effects					
Bullying -->School Violence	-0.538	0.036	-0.302	-0.755	0.000
School Violence-->Depression/Suicidality	-0.476	0.021	-0.490	-0.490	0.000
Total effects	0.495	0.025	0.285	0.715	0.000
Total Indirect	0.256	0.018	0.148	0.370	0.000

The unstandardized model parameters assume all variables are on the same scale, which is not the case here. Therefore, we have to use standardized parameters to describe the relationship between the variables. Mplus provides three types of standardization options: STDYX, STDY, and STD. All the direct and indirect effects were significant.. STDYX says both the dependent and independent variables are standardized, while STD only standardizes the dependent variable. In general, StdYX and Std are the same when the parameter estimates only involve relationships between latent variables. However, when a factor is regressed on a binary observed variable then STDY or STD is used.

From the Mplus output, there is a significant direct effect (β=.239, *SE*=.023, p < .001) and indirect effect (β=.256, *SE*=.018, p < .001) from the bullying variable to the variable on depression and suicidality. Looking at the standardized direct effect those who experience no bullying had a .345 standard deviation increase on the depression/suicidality scale than those who had experienced bullying, controlling for the effect of school violence. The latent depression/suicidality scale goes in the direction of the first three indicators from yes, or affirmative depression/suicidality, to negative depression suicidality. Therefore, those who have not experienced bullying have experience less depression/suicidality while those who have experienced bullying have a .345 standard deviation decrease (experience more) depression/suicidality. The indirect effect operates through school violence, so according to the standardized (STD) estimates, those who say they have not been bullied have a decreased of .755 standard deviation on school violence scale compared to those who have been bullied. Since the school violence scale goes from no school violence to extreme amount of school violence, this means those who have been bullied experience more school violence. In turn, the one standard deviation increase on the school violence scale leads to .490 standard deviation decrease on the depression and suicidality scale. Since this scale goes from more depression/suicidality to less, this decrease indicates higher levels of depression/suicidality. The total mediated effect, c= c' +a*b, is significant (β=.495, *SE*=.025, p<.001). Using unstandardized parameters, we see that .495=.239 + .256 (where .256= -.476*-.538). So, if we were not controlling for the mediated variable we would see for those who said they have been not bullied saw a .495 increase in depression and suicidality compared to those who said they had been bullied (recall, the parameters are not standardized), or rather those who haven't experience any bullying experience less depression/suicidality.

Figure 1.9

Teen Alcohol Abuse and Usage Mediation Model



The second model modeled the relationship between school violence and depression/suicidality mediated by teen alcohol abuse and usage. The weighted least square mean and variance (WLSMV) estimator was used and 16409 observations were used in the analysis using pairwise deletion methodology. The RMSEA=.030 (90% CI: .028 .032), CFI= .987, TLI=.982, and WRMR =2.731. The weighted root mean square residual (WRMR) indicates poor fit but all the other indices suggest good fit. The Modification indices (MI) indicate parameters that can be added to the model to improve fit by reducing χ^2. The biggest reduction of chi-square was the suggested cross loading of drinking at school onto school violence (440.312). Next was the cross loading of drinking at school (DRKSCH) onto depression and suicidality (236.045). The third largest was loading weapons at school (WSCH) onto teen alcohol abuse and usage (106.261). Although these sound interesting, they have no theoretical basis so I could not incorporate the suggested modifications. Another concern was the high, almost perfect loadings. According to Kline (2005), when a correlation is greater than .90, we have redundant information. Drinking at least one serving of alcohol in the past 30 days was the second most missed variable and it had a high correlation with drinking at least one alcohol in one's life. They are basically saying the same thing so the model was refitted without this variable. After removing this variable, the new model fit indices are as follows: RMSEA=.031 (90% CI: .029 .033), CFI= .968, TLI=.959, and WRMR =2.695. This was a modest improvement but still necessary based on the high correlation. Figure 1.10 outlines the new model with the new parameter estimates. Table 1.3 features the parameter estimates for the mediation model. The loadings were all significant. The explained variance ranged from .292 to .908.

Figure 1.10

Corrected Teen Alcohol Abuse and Usage Mediation Model

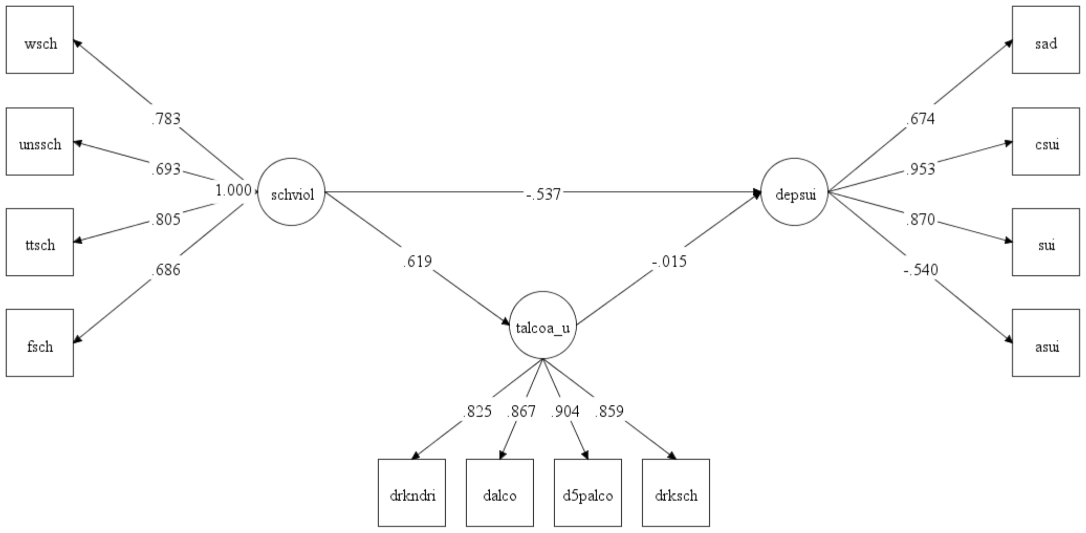
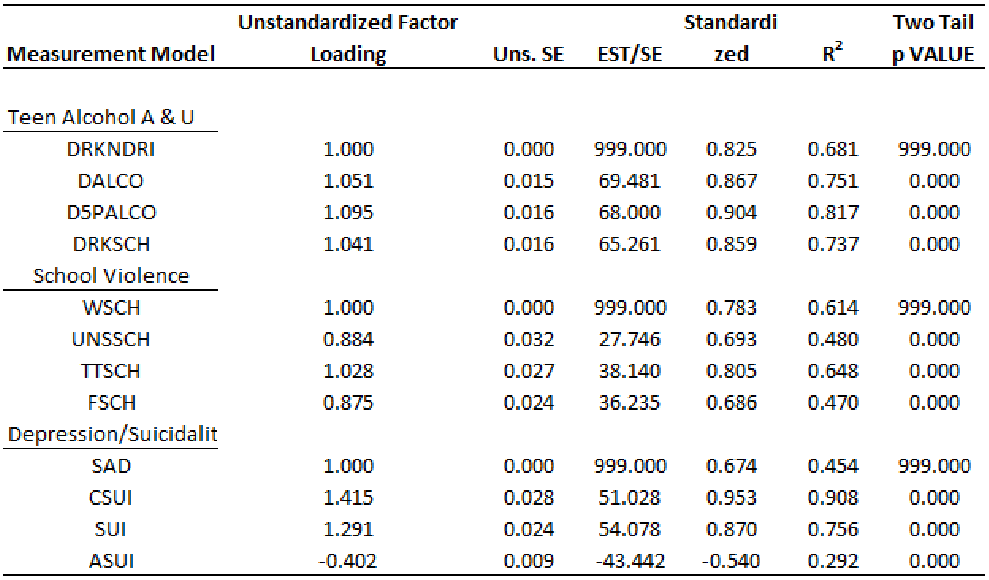


Table 1.3

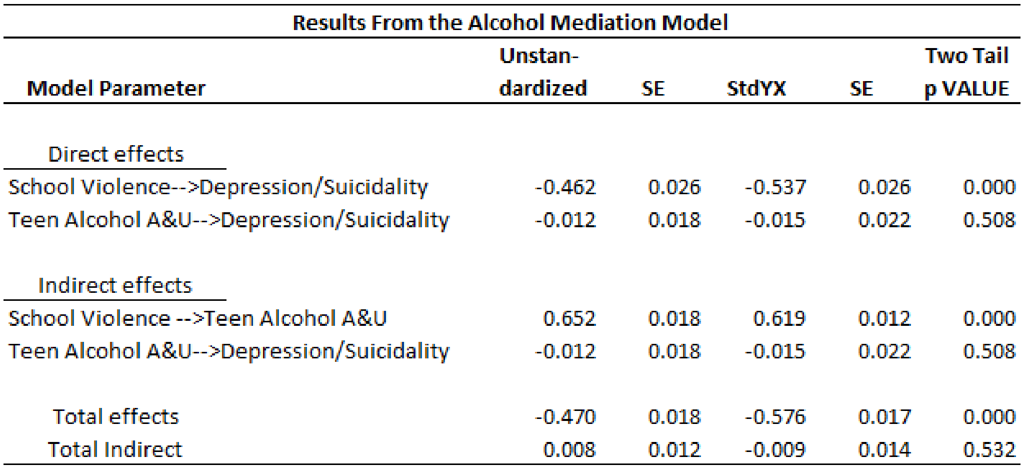
Measurement Model Alcohol Mediation



	Unstandardized Factor			Standardi		Two Tail
Measurement Model	Loading	Uns. SE	EST/SE	zed	R^2	p VALUE
Teen Alcohol A & U						
DRKNDRI	1.000	0.000	999.000	0.825	0.681	999.000
DALCO	1.051	0.015	69.481	0.867	0.751	0.000
D5PALCO	1.095	0.016	68.000	0.904	0.817	0.000
DRKSCH	1.041	0.016	65.261	0.859	0.737	0.000
School Violence						
WSCH	1.000	0.000	999.000	0.783	0.614	999.000
UNSSCH	0.884	0.032	27.746	0.693	0.480	0.000
TTSCH	1.028	0.027	38.140	0.805	0.648	0.000
FSCH	0.875	0.024	36.235	0.686	0.470	0.000
Depression/Suicidalit						
SAD	1.000	0.000	999.000	0.674	0.454	999.000
CSUI	1.415	0.028	51.028	0.953	0.908	0.000
SUI	1.291	0.024	54.078	0.870	0.756	0.000
ASUI	-0.402	0.009	-43.442	-0.540	0.292	0.000

Table 1.4

Direct/Indirect Effects Alcohol Mediated Model



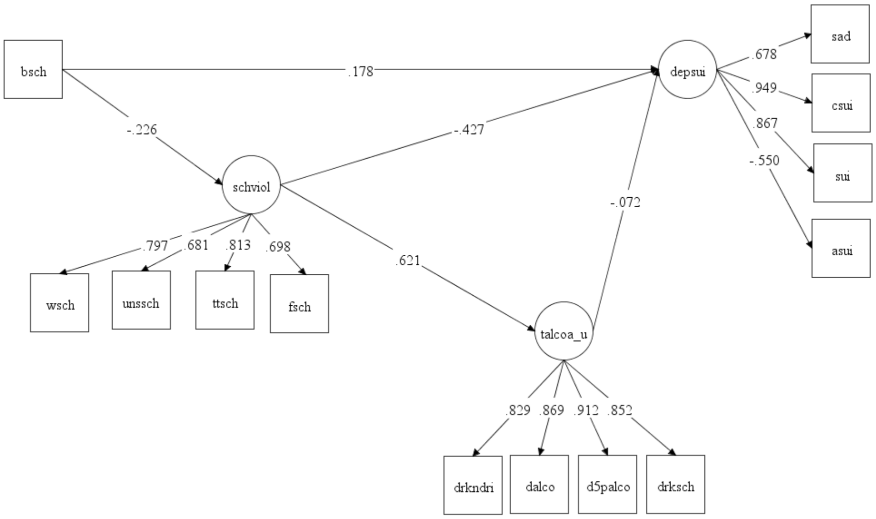
Results From the Alcohol Mediation Model					
	Unstan-				Two Tail
Model Parameter	dardized	SE	StdYX	SE	p VALUE
Direct effects					
School Violence-->Depression/Suicidality	-0.462	0.026	-0.537	0.026	0.000
Teen Alcohol A&U-->Depression/Suicidality	-0.012	0.018	-0.015	0.022	0.508
Indirect effects					
School Violence -->Teen Alcohol A&U	0.652	0.018	0.619	0.012	0.000
Teen Alcohol A&U-->Depression/Suicidality	-0.012	0.018	-0.015	0.022	0.508
Total effects	-0.470	0.018	-0.576	0.017	0.000
Total Indirect	0.008	0.012	-0.009	0.014	0.532

According to Table 1.4, school violence has a significant direct effect on depression/suicidality (β_{StdXY} = -.537, *SE*=.026, p < .001). Each standard deviation increase on the school violence scale leads to a .537 standard deviation decrease on the depression/suicidality scale. Since the depression/suicidality scale go from very depressed and suicidal to zero, a decrease means the more violent the school the more students experienced depression and suicidality. Although there is a significant relationship between school violence and alcohol use and abuse (β_{StdXY} = -.619, *SE*=.012, p < .001), there is not a significant relationship between depression/suicidality and alcohol abuse and use (β_{StdXY} = -.015, *SE*=.022, p =.508). Thus, the alcohol abuse and use is not a mediator between school violence and depression/suicidality.

Model 3 has dual or competing mediators, school violence and teen alcohol abuse and abuse, see figure 1.11. Examination of model fit reveals that RMSEA=.030 (90% CI: .028 .032), CFI= .965, TLI=.955, and WRMR =2.748. There is evidence of model fit using RMSEA, CFI, and TLI. WRMR does not show that the model fit.

Figure 1.11

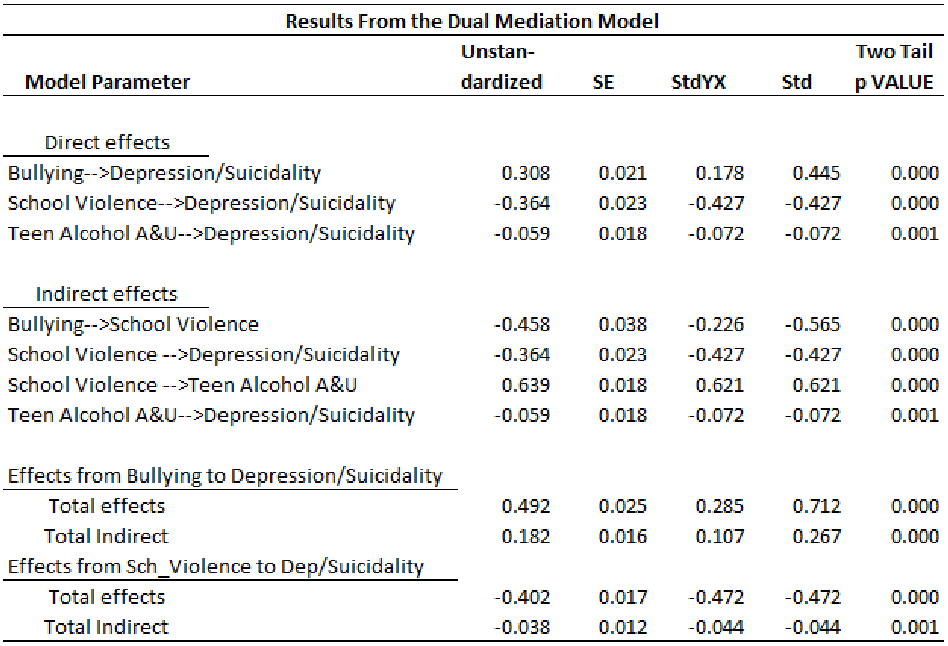
Dual Mediator model



Looking at the direct and indirect effects for the dual mediation model (see table 1.5) I found that all paths were significant. The direct effect of school violence on depression/suicidality is significant and negative (β= -.364, *SE*=.023, p < .001). According to the standardized (STDXY) estimates, for each standard deviation increase on the school violence scale there is .364 standard deviation decreases on the depression/suicidality scale. In addition, *school violence* had a significant indirect effect that was mediated by *teen alcohol abuse & use*, β = -.038, *SE*=.012, p = .001. Each standard deviation increase in school violence leads to a .044 (.621*-.072) standard deviation decrease in depression/suicidality because school violence operates through teen alcohol abuse and use.

Table 1.5

Direct/Indirect Effects Dual Mediated Model



Results From the Dual Mediation Model

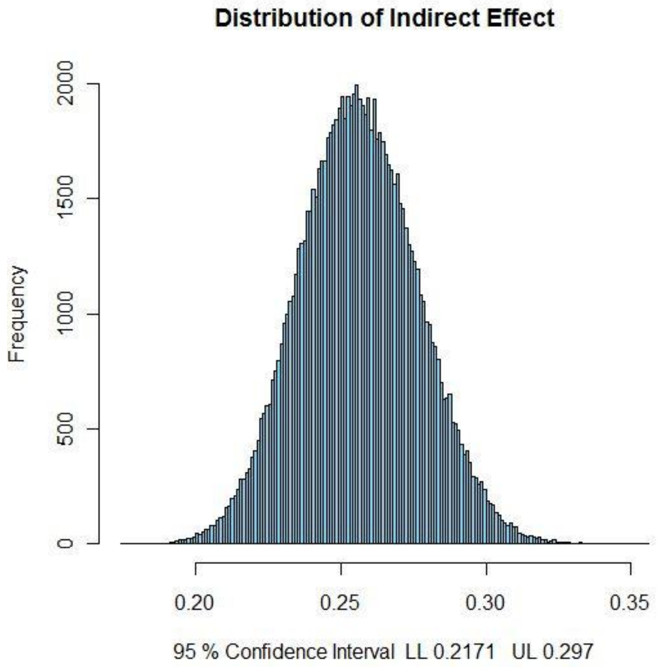
Model Parameter	Unstan- dardized	SE	StdYX	Std	Two Tail p VALUE
Direct effects					
Bullying-->Depression/Suicidality	0.308	0.021	0.178	0.445	0.000
School Violence-->Depression/Suicidality	-0.364	0.023	-0.427	-0.427	0.000
Teen Alcohol A&U-->Depression/Suicidality	-0.059	0.018	-0.072	-0.072	0.001
Indirect effects					
Bullying-->School Violence	-0.458	0.038	-0.226	-0.565	0.000
School Violence -->Depression/Suicidality	-0.364	0.023	-0.427	-0.427	0.000
School Violence -->Teen Alcohol A&U	0.639	0.018	0.621	0.621	0.000
Teen Alcohol A&U-->Depression/Suicidality	-0.059	0.018	-0.072	-0.072	0.001
Effects from Bullying to Depression/Suicidality					
Total effects	0.492	0.025	0.285	0.712	0.000
Total Indirect	0.182	0.016	0.107	0.267	0.000
Effects from Sch_Violence to Dep/Suicidality					
Total effects	-0.402	0.017	-0.472	-0.472	0.000
Total Indirect	-0.038	0.012	-0.044	-0.044	0.001

Bullying has a significant direct effect on depression/suicidality as well, β= .308, *SE*=.021, p < .001. According to the standardized (STD) estimates, the absence of bullying leads to a .445 standard deviation increase on the depression/suicidality scale compared to those who said they had been bullied, controlling for effects of school violence, and alcohol abuse and use in the model. There are two indirect paths through which bullying can influence depression/suicidality. The first is through school violence (Bullying --> School violence --> Depression/Suicidality), which is significant (β= -.364, *SE*=.023, p < .001). Using the standardized parameters (STD) to interpret the relationship, those who have not been bullied will see an increase by .242 (-.565*-.427) standard deviations through this path compared to those who have been bullied. The second path is through school violence and then through teen alcohol abuse and use (Bullying --> School Violence -->Teen Alcohol Abuse & Use --> Depression/Suicidality); this path is also significant (β= -.017, *SE*=.005, p = .001). Those who were not bullied will see an increase of .025 (-.565* .621*-.072) standard deviations on the depression/suicidality scale mediated through school violence and teen alcohol abuse use. The total effect of bullying through all three pathways (one direct and two indirect) is an increase on the depression/suicidality scale by .712 (.445 + .241+.025) standard deviations for those who experienced no bullying versus those who experienced bullying. Explicitly, those who experience bullying experience a great deal more depression/sucidality.

The test for indirect or mediated effect, (i.e., a*b is significantly different from zero) is by default the Sobel test in Mplus (Geiser, 2013). The problem with using the Sobel test to test for indirect effect is that it assumes the mediation paths, a*b, is normally distributed, which is rarely the case. Instead, it is better to look at an asymmetric confidence interval based on bootstrap methodology to test for indirect effects. However, bootstrapping can't be run on multilevel or complex data in Mplus. One possible good note, the problem of the distribution of a*b not being normal comes when the sample size is small; for moderate to large sample sizes the distribution may approach normality (Bollen & Stine, 1990). Although bootstrapping, a non-parametric technique, is recommended when non-normality is present, Monte Carlo confidence intervals (MCCI) for indirect effects is another approach that is just as effective and not as time consuming as bootstrapping (Preacher & Selig, 2012). This approach generates asymmetric confidence intervals as well and can be used in placed of bootstrapping. Preacher and Selig (2012) provided a general R script to run MCCI for indirect effect but the code has to be adapted for your particular model. Using his starter code written in R, I was able to transfer information provided under tech1 and tech3 in Mplus (location of covariance matrix for the paths used in the calculation for the indirect effect) and implement their code (see Appendix B).

Figure 1.12

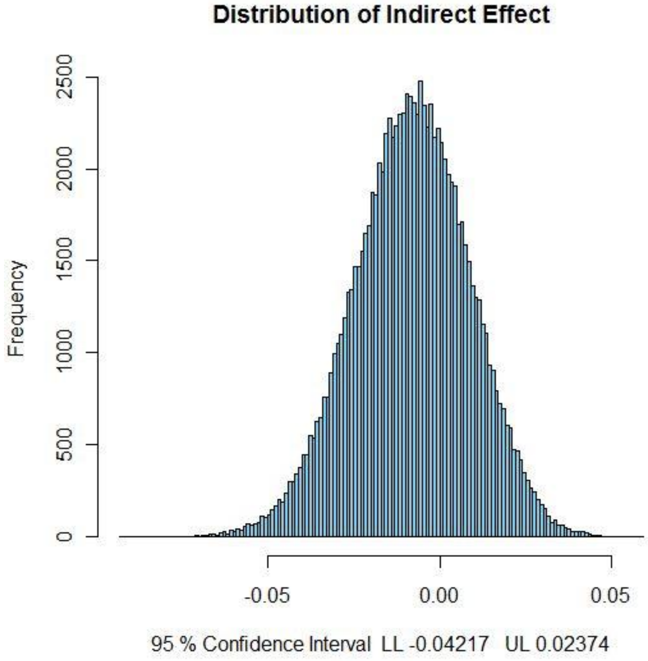
MCCI for Indirect effect Model 1



The monte carlo confidence interval for the total indirect effect of model one, where *school violence* mediated the relationship between *bullying* and *depression/suicidality*, was shown (or confirmed) to be significant with the confidence interval of (.2171, .297). The unstandardized total mediation effect was .256. Note, the unstandardized values were used in the R code.

Figure 1.13

MCCI for Indirect effect Model 2



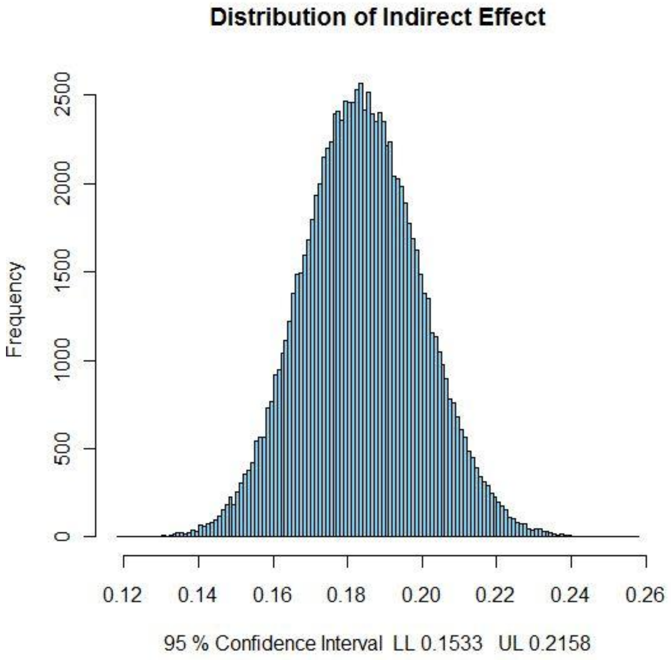
The MCCI for the second mediation model where *teen alcohol abuse and use* was tested to mediate the relationship between *school violence* and *depression/suicidality* was confirmed to not be significantly different from zero (MCCI = -.04217, .02374).

The MCCI for the full dual mediation model, model3, see figure 14, was shown to be significant with a 95% CI bounded away [.1522, .2158]. The unstandardized total indirect effect of this model was .182.

Using Preacher and Selig's (2012) algorithm for measuring if there is an indirect effect for asymmetric distributions, I was able to confirm the same results seen using Sobel significance test.

Figure 1.14

MCCI for Indirect effect Model 3



CHAPTER FIVE

STUDY ONE DISCUSSION

This section discusses the findings in the results section. There were three questions I wanted to answer in this research: Is there a relationship between bullying and depression/suicidality? Is that relationship mediated by school violence? Is teen alcohol abuse and use another possible mediator in the model? In addition, I want to know if there possible other mediators for this model? I hypothesized that there would be a relationship between bullying and depression/suicidality and it would be mediated by school violence. I hypothesized that teen alcohol abuse and use would mediate the relationship between school violence and bullying.

The empirical data reveal that there is a relationship between bullying and depression/suicidality. Recall, the bullying scale went from one to two where one was the affirmative answer to having not experienced bulling during the past 12 months. So being bullied was the baseline or reference group. The depression/suicidality scale had a reverse order with those having a positive standard deviation increase on the scale was actually experiencing less depression/suicidality.

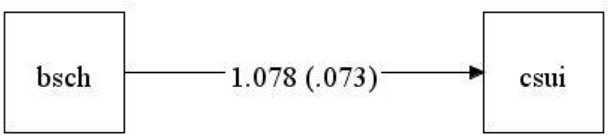
The relationship between bullying and depression/sucidality was mediated by school violence. As a standalone model, teen alcohol abuse and use did not mediate the relationship between school violence and depression/suicidality. However, when added as a second mediator to the first model, teen alcohol abuse and use was a mediator between school violence and depression/suicidality, and it was a secondary mediator between bullying and depression/suicidality. The total indirect mediation was significant probably because the

Are there possible other mediators in this model? Yes! According to Zhao et al. (2010), when there is a significant indirect effect, significant total direct effect, and a*b*c is positive then this provides evidence of an omitted potential mediator and the model has an incomplete framework.

Another question that could reasonably be asked, Is SEM the model to use for this data set? The CDC advises researchers to use the logistic regression model and correct for the standard errors. I looked at a particular logistic regression model, and I regressed the observed variable considering suicide on the observed bullying variable using Mplus 7.0. *Considered suicide* is a dichotomous variable with one for yes and two for no. The results showed that there was a significant relationship between bullying and depression/suicidality (β= 1.078, *SE*=.073, p < .001), Figure 1.15. Those who have not experienced bullying add an increased odd of 1.078 of not considering suicide (my reference variable is yes considering suicide).

Figure 1.15

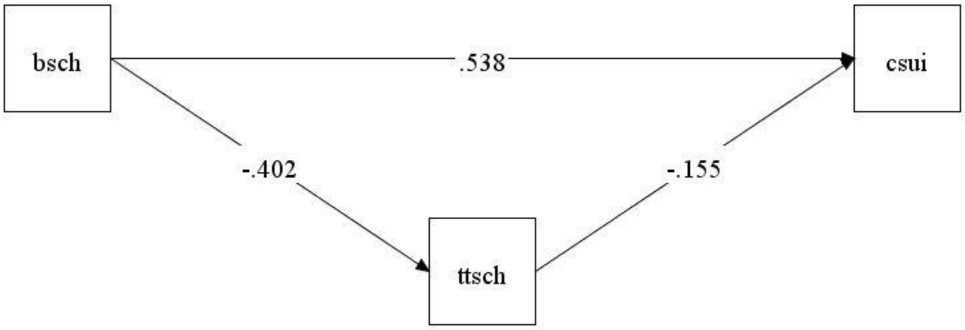
Logistic Regression Model



When *threatened at school* is added to the model as a mediator, figure 1.16, we see all paths are significant (β=0.538, *SE*=.040, p < .001). There is a .538 increase in odds of not considering suicide for those who have not experience bullying. There was a significant indirect effect as well (β=0.063, *SE*=.006, p < .001).

Figure 1.16

Indirect effect for observed model



Looking at the R-square, or percentage of variation explained by independent variables in the model with the observed mediated model (figure 1.16), only 7.4% of the variance in *considered suicide* was explained by *threated at school* and *bullying*. However, for model one, 30% of the variance in *depression/suicidality* was explained by the model. Although, more variance was explained by the SEM model, both models still have the same handicap, it requires students to define what is bullying to them and require them to remember.

The YRBS is a useful tool and great data set but is limited in that the full bullying construct has not been explored as well as other constructs in the model.

APPENDIX A

TABLE A.1 SCALE

Variables/Indicators	**Scale**
Bully	
During the past 12 months, have you ever been *bullied on school property? (BSch)*	A.Yes B. No
School Violence	**Scale**
During the past 30 days, on how many days did you *carry a weapon such as a gun, knife, or club on school property? (WSch)*	A. 0 days B. 1 day C. 2 or 3 days D. 4 or 5 days E. 6 or more days
During the past 12 months, how many times were you in a *physical fight on a school property? (FSch)*	A. 0 times B. 1 time C. 2 or 3 times D. 4 or 5 times E. 6 or 7 times F. 8 or 9 times G. 10 or 11 times H. 12 or more times
During the past 30 days, on how many days did you not go to school because you *felt you would be unsafe at school or on your way to or from school? (UnSSch)*	A. 0 days B. 1 day C. 2 or 3 days D. 4 or 5 days E. 6 or more days
During the past 12 months, how many times has *someone threatened or injured you with a weapon such as a gun, knife, or club on school property? (TtSch)*	A. 0 times B. 1 time C. 2 or 3 times D. 4 or 5 times E. 6 or 7 times F. 8 or 9 times G. 10 or 11 times H. 12 or more times

Suicidality	**Scale**
During the past 12 months, did you ever feel so *sad or hopeless* almost every day for two weeks or more in a row that you stopped doing some usual activities? *(Sad)*	A.Yes B. No
During the past 12 months, did you ever seriously *consider attempting suicide*?*(CSui)*	A.Yes B. No
During the past 12 months, did you make a *plan about how you would attempt suicide?(Sui)*	A.Yes B. No
During the past 12 months, how many times did you *actually attempt suicide?(Asui)*	A. 0 times B. 1 time C. 2 or 3 times D. 4 or 5 times E. 6 or more times
Teen Alcohol Abuse and Use	**Scale**
During the past 30 days, how many times did you *drive a car or other vehicle when you had been drinking alcohol?(DrkNDri)*	A. 0 days B. 1 day C. 2 or 3 days D. 4 or 5 days E. 6 or more days
During your life, on how many days have you had at *least one drink of alcohol?(DAlco)*	A. 0 days B. 1 or 2 days C. 3 to 9 days D. 10 to 19 days E. 20 to 39 days F. 40 to 99 days G. 100 or more days
During the *past 30 days*, on how many days did you have at *least one drink of alcohol*? *(DL1Alco*)	A. 0 days B.1 to 2 days C. 3 to 5 days D. 6 to 9 days E. 10 to 19 days F. 20 to 29 days G. All 30 days
During the past 30 days, on how many days did you have *5 or more drinks of alcohol in a row,* that is, within a couple of hours?*(D5PAlco*)	A. 0 days B. 1 day C. 2 days D. 3 to 5 days E. 6 to 9 days F. 10 to 19 days G. 20 or more days

During the past 30 days, on how many days did you have at *least one drink of alcohol on school property?(DrkSch)*	A. 0 days B.1 to 2 days C. 3 to 5 days D. 6 to 9 days E. 10 to 19 days F. 20 to 29 days G. All 30 days

TABLE A.2

DISTRIBUTION

Variables/Indicators	Distribution
Bully	
During the past 12 months, have you ever been *bullied on school property? (BSch)*	A. 18.9 B. 81.1
School Violence	
During the past 30 days, on how many days did you *carry a weapon such as a gun, knife, or club on school property? (WSch)*	A. 94.5 B. 1.5 C. 1.1 D..3% E. 2.5%
During the past 12 months, how many times were you in a *physical fight on a school property? (FSch)*	A. 89.3 B. 6.9 C. 2.5 D. .4 E. .3% F. .1 G. 0 H. .6%
During the past 30 days, on how many days did you not go to school because you *felt you would be unsafe at school or on your way to or from school? (UnSSch)*	A. 95.8 B. 2.1 C. 1.1 D. 0.3 E. .7
During the past 12 months, how many times has *someone threatened or injured you with a weaspon such as a gun, knife, or club on school property? (TtSch)*	A. 93.2 B. 3.1 C. 1.7 D. .6 E. .3 F. .2 G. .1 H. .9
Suicidality	
During the past 12 months, did you ever feel so *sad or hopeless* almost every day for two weeks or more in a row that you stopped doing some usual activities? *(Sad)*	A. 27.1 B. 72.9
During the past 12 months, did you ever seriously *consider attempting suicide*?*(CSui)*	A. 14.1 B. 85.9

During the past 12 months, did you make a *plan about how you would attempt suicide?(Sui)*	A. 11.2% B. 88.8%
During the past 12 months, how many times did you *actually attempt suicide?(Asui)*	A. 93.5 B. 3.5 C. 1.9% D. .4% E. .7%
Teen Alcohol Abuse and Use	
During the past 30 days, how many times did you *drive a car or other vehicle when you had been drinking alcohol?(DrkNDri)*	A. 89.7% B. 4.6% C. 3.4% D. .8% E. 1.5%
During your life, on how many days have you had at *least one drink of alcohol?(DAlco)*	A. 26.3% B. 16.5% C.17.1 D. 11.6% E. 10.1% F. 8.6% G. 9.7%
During the *past 30 days*, on how many days did you have at *least one drink of alcohol*? *(DL1Alco*)	A. 57.2% B. 21.3% C. 10.6% D. 5.6% E. 3.8% F. .6% G. .9%
Druing the past 30 days, on how many days did you have *5 or more drinks of alcohol in a row,* that is, within a couple of hours?*(D5PAlco*)	A. 74.2% B. 9.0% C. 6.3% D. 5.8% E. 2.7% F. 1.3% G. .8%
During the past 30 days, on how many days did you have at *least one drink of alcohol on school property?(DrkSch)*	A. 95.1% B.3.2% C. .9% D. .2% E. .1% F. 0% G. .4%

TABLE A.3

CORRELATION

	DRKNDRI	DALCO	DL1ALCO	D5PALCO	DRKSCH	BSCH	WSCH	UNSSCH	TTSCH	FSCH	SAD	CSUI	SUI
DRKNDRI													
DALCO	0.704												
DL1ALCO	0.779	0.846											
D5PALCO	0.747	0.816	0.926										
DRKSCH	0.289	0.335	0.405	0.362									
BSCH	-0.060	-0.085	-0.077	-0.075	-0.082								
WSCH	0.534	0.459	0.480	0.449	0.247	-0.171							
UNSSCH	0.366	0.216	0.310	0.280	0.193	-0.388	0.458						
TTSCH	0.473	0.343	0.394	0.388	0.239	-0.430	0.567	0.620					
FSCH	0.392	0.294	0.351	0.333	0.258	-0.258	0.528	0.469	0.579				
SAD	-0.160	-0.225	-0.222	-0.185	-0.110	0.345	-0.190	-0.388	-0.341	-0.231			
CSUI	-0.193	-0.236	-0.216	-0.199	-0.135	0.343	-0.298	-0.380	-0.355	-0.276	0.660		
SUI	-0.218	-0.249	-0.230	-0.195	-0.140	0.358	-0.325	-0.384	-0.383	-0.279	0.601	0.858	
ASUI	0.196	0.172	0.206	0.178	0.411	-0.176	0.224	0.234	0.264	0.228	-0.333	-0.513	-0.448

TABLE A.4 MISSINGNESS

Variables/Indicators	Missingness
Bully	
During the past 12 months, have you ever been *bullied on school property? (BSch)*	777
School Violence	
During the past 30 days, on how many days did you *carry a weapon such as a gun, knife, or club on school property? (WSch)*	154
During the past 12 months, how many times were you in a *physical fight on a school property? (FSch)*	321
During the past 30 days, on how many days did you not go to school because you *felt you would be unsafe at school or on your way to or from school? (UnSSch)*	39
During the past 12 months, how many times has *someone threatened or injured you with a weapon such as a gun, knife, or club on school property? (TrSch)*	43
Suicidality	
During the past 12 months, did you ever feel so *sad or hopeless* almost every day for two weeks or more in a row that you stopped doing some usual activities? *(Sad)*	178
During the past 12 months, did you ever seriously *consider attempting suicide*?*(CSui)*	190
During the past 12 months, did you make a *plan about how you would attempt suicide?(Sui)*	197
During the past 12 months, how many times did you *actually attempt suicide?(Asui)*	1801

Teen Alcohol Abuse and Use	
During the past 30 days, how many times did you *drive a car or other vehicle when you had been drinking alcohol?(DrkNDri)*	289
During your life, on how many days have you had at *least one drink of alcohol?(DAlco)*	457
During the *past 30 days*, on how many days did you have at *least one drink of alcohol*? *(DL1Alco*)	1546
During the past 30 days, on how many days did you have *5 or more drinks of alcohol in a row,* that is, within a couple of hours?*(D5PAlco*)	401
During the past 30 days, on how many days did you have at *least one drink of alcohol on school property?(DrkSch)*	386

APPENDIX B

Preacher & Selig (2012) Monte Carlo Indirect effect intervals code

Model3

```
require(MASS)

a1 <- -.458

a2 <- 0.639

b1 <- -.364

b2 <- -0.059

rep=100000

conf=95

pest <- c(a1,a2,b1,b2)

acov <- matrix(c(


 .0014165726, 0.00023418028, -0.000011796112, -0.000099516474,

 0, 0.00031608746, -0.0000029708807, -0.000064930249,

 0, 0, 0.00052616222, -0.00030901698,

 0, 0, 0, 0.00033035410


),4,4)

mcmc <- mvrnorm(rep,pest,acov,empirical=FALSE)
```

```
ie <- mcmc[,1]*mcmc[,3]+mcmc[,1]*mcmc[,2]*mcmc[,4]
low=(1-conf/100)/2
upp=((1-conf/100)/2)+(conf/100)
LL=quantile(ie,low)
UL=quantile(ie,upp)
LL4=format(LL,digits=4)
UL4=format(UL,digits=4)
print(c(a1*b1+a1*a2*b2,LL,UL))
hist(ie,breaks='FD',col='skyblue',xlab=paste(conf,'% Confidence Interval ','LL',LL4,' UL',UL4),main='Distribution of Indirect Effect')
```

SECTION TWO

SAMPLE SIZE REQUIREMENT FOR
NON-NORMAL COMPLEX MULTILEVEL
DATA FOR THE MULTILEVEL STRUCTURAL EQUATION MODEL

CHAPTE ONE

STUDIES TWO AND THREE INTRODUCTION

Data in real life are rarely perfect. When learning statistics, you often learn the four basic assumptions that underlie the general linear model: the assumptions of linearity, independence, constant variance, and normality. When the relationship between the dependent and independent variable is non-linear or there is non-constant variance about the regression line, then there are often tools one can use to remedy this situation, such as transformations. The assumption violated and the severity of the violation dictates which tool or statistical technique to use. Some violations are often tied to disciplines/subjects.

In educational and psychological research, data are often non-experimental data, such as survey data. These types of data disallow experiments where one can manipulate the variables to test cause and effect. For this type of data (non-experimental), Structural Equation Modeling (SEM) is often advocated and used to adjust for the nature of this limitation (Bryne, 2012). Although pervasive in educational and psychological research, SEM is rarely used in other disciplines such as medical research (Kupek, 2006). Another known problem with educational and psychological data are that the data are rarely normal. Micceri (1989) examined 440 data sets from educational and psychological research where half of those data sets were used in published journal articles, and found normality was non-existent. In fact, non-normality was the rule rather than the exception. Structural equation modeling can accommodate non-normality and non-experimental data. However, one of the main assumptions that underlie SEM and general (or generalized) linear models is independence. Violation of independence is a very serious violation of statistical design. When observations are not independent, then there is an increased risk of type I error which can lead to incorrect conclusions due to the underestimation of the true variability (Keppel, 2004; Heck & Thomas, 2000). The data from complex samples are often dependent due to cost saving techniques employed during data collection. Multilevel modeling (or hierarchical modeling) was introduced to address the issue of dependent data. Multilevel modeling (MLM) has many synonyms that are discipline specific, it is called mixed modeling by the department of statistics, hierarchical linear modeling (HLM) by educational researchers, and also random effects modeling. When combined with SEM, we have multilevel structural equation modeling (MSEM). Performing HLM or MSEM one can correct the negative standard error bias inherit when analysis is run ignoring the hierarchical in nature of the data. MSEM has an added benefit in that it allows for the modeling of the structure of the relationship at various levels of the model. In general, being able to test several plausible models and including measurement error in the model, which leads to less bias parameter estimates when variables are assumed to be measured without error, are some of the benefits that make using SEM on non-experimental data popular. MSEM adopts these same benefits, as well as the ability for the dependent variable to not strictly be at the lower level.

Purpose & Significance

The idea for this study surfaced from complication of running MSEM on a national data set where the use of SEM was seldom used or advised to be used. In researching the reason for the computer convergence problems, I found that very little research has been done on non-normal multilevel SEM in regards to sample size requirement which is one of the reasons for non-convergence. Hox and Maas (2010) found that sample size recommendations depended on the estimation method. However, their recommendations and studies were done for continuous multivariate normal data. They stated that no one had studied sample size recommendations for non-normal data within a multilevel SEM context (Hox & Maas, 2010). Also, Preacher (2011) stated that sample size recommendations for non-normal data have not been studied. Since data in educational research is often non-normal, and data within schools have inherent dependencies, studying sample size requirement under non-normality was needed. This study combines two studies: 1) studying sample size requirements when data are non-normal continuous, 2) studying sample size when data are categorical. The purpose is to shed light on a much needed area of research, and to give guidelines to future researches on the minimum requirement needed to run MSEM on non-normal complex data. This study is significant because it has not been done before.

Research Questions

Study two is a Monte Carlo study that looks at the sample size requirement when the data are non-normal continuous. There are three research questions that I seek to answer:

Does sample size requirement for non-normal continuous data depend on the estimation method?

Is the sample size requirement greater for normal or non-normal continuous data for the respective estimation method?

Does the presence or absence of unbalanced clusters affect the sample size requirement for non-normal continuous data?

Study three is a Monte Carlo study that looks at sample the size requirement when the data are categorical. There are four research questions that I would like to answer in this study:

Does sample size requirement for categorical independent variable data depend on estimation method?

Is the sample size requirement the same or different compared to the normal multilevel data for the respective estimation method?

Does the presence or absence of unbalanced clusters affect the sample size requirement for categorical data?

Does the presence of sparse tables affect the sample size requirement?

CHAPTER TWO

STUDIES TWO AND THREE REVIEW OF LITERATURE

Studies two and three arose from the need to model multilevel structural equation modeling on non-normal non-experimental data from a national data set. Because there wasn't any guideline in the literature detailing what sample size was needed under non-normal multilevel SEM data, studies two and three sought to address this gap in the literature. Study two seeks to find the sample size requirement when we have continuous non-normal multi-level data, while study three seeks to find the sample size requirement when data are categorical multi-level data. This literature review for the two studies is divided into seven distinct sections. The first section explores the utility of SEM when non-experimental data are present. A careful presentation of the advantages and limitations of using SEM is presented when we have non-experimental data. Non-experimental data consist of observational, survey, and data from an existing data set (such as a national data set). It is very common in educational research to have non-experimental data and for SEM to be used in modeling this non-experimental data (Muijs, 2011; Bryne, 2012).Then we have a brief introduction to structural equation modeling, and a brief introduction to multi-group structural equation modeling. Lastly, an introduction to Multi-indicator Multi-independent causes Models (MIMIC), multilevel SEM (MSEM), mixture models, and estimations in SEM is included. Within these topics a deeper understanding of the limitations of SEM, the relationship between MSEM and multi-group SEM and MIMIC, and how heterogeneity is introduced into a distribution is sought.

Problem with non-experimental data and advantages of SEM

The strongest study is a study that can make statements about cause and effect (Berg & Latin, 2004). There are three conditions for establishing a causal relationship: temporal precedence (i.e., independent variable precedes the dependent variable in time), co-variation (i.e., the independent variable and dependent variable co-vary, meaning when one changes the other changes), and all other possible explanations/variables have been ruled out (i.e., no confounding variables; Johnson & Kruse, 2009). Research design can be experimental or non-experimental. Experimental studies are considered the gold standard of research because cause and effect can be ascertained. Experimental studies establish relationships by manipulating the independent variables and then observing the outcome, thus keeping the time order requirement. By utilizing the randomization process, other possible explanations are ruled out. Not all variables can be manipulated such as income, GPA, or self-esteem. You can't force someone to make less income and see what happens to the dependent variable. With non-experimental studies the temporal order is out of order; data are collected for the dependent and independent variable at the same time. The dependent and independent variables co-vary but other possible explanations cannot be ruled out. With non-experimental research designs it is impractical, if not impossible, to manipulate some independent variables. So instead of an experiment other research designs are used such as surveys, observational studies, or correlational studies.

The problem with linear regression is that the independent variable is assumed to be measured without error. However, when measurement error is present the regression equations changes. Instead of there being of relation between the true independent variable and dependent variable, the relationship is between a surrogate (what was measured) variable and the independent variable.

$$X^* = X + E \tag{7}$$

The true independent variable is X but we are measuring X*. The linear regression equation becomes

$$Y = \beta X^* + \varepsilon \tag{8}$$

instead of the true model (intercept ignored) $Y = \beta X + \varepsilon$. This new equation adds extra components (error) to the model. Rewriting the equation in terms of X*, X=X*-E,

$$\begin{aligned} Y &= \beta(X^* - E) + \varepsilon \\ Y &= \beta X^* + (\varepsilon - \beta X^*) \end{aligned} \tag{9}$$

We can see a new error term has been included introduced in the model. So when Y is regressed on X* the error term is, $\varepsilon - \beta X^*$. This means that the independent variable and the error term are not independent, which is an assumption in linear regression (Graddy & Wang, 2008). Also according to Carroll, Ruppert, and Stefanski (2006), the beta coefficient is affected by the presence of measurement error. The reliability ratio, λ, gives the measure of how much attenuation occurs in the explanatory variable.

$$\lambda = \frac{\sigma_x^{\ 2}}{\sigma_x^{\ 2} + \sigma_\varepsilon^{\ 2}} \tag{10}$$

When we are modeling Y on X, the actual equal is not beta but a function of beta, β*= λβ. If X is perfectly reliable (lambda equals one) then the new beta and true beta are equal. So for our model with measurement error the model equation becomes

$$\begin{gathered} Y = \beta * X * + (\varepsilon - \beta * X*) \\ Y = (\lambda\beta)X * + (\varepsilon - \lambda\beta X*) \end{gathered} \tag{11}$$

$$Var(Y) = \lambda^{2}\beta^{2}Var(X*) + (Var(\varepsilon) + \lambda^{2}\beta^{2}Var(X*)) \tag{12}$$

The addition of measurement error causes additional error variance in the model. If you look at the scatter plot with the regression line, you will see more error about the line. This error according to Carroll et. al. (2006) attenuates the slope. If the measurement error is small and random then one can ignore it, otherwise, an alternative estimation strategy needs to be explored (Graddy & Wang, 2008).

Researchers are told that when you have non-experimental data (that has measurement error) that the best tool to use is structural equation modeling (SEM). Fabrigar, Porter, Norris (2010) noted that modeling causal relations was not the same as providing a causal relationship or conclusion. It is not the statistical tool used that causes the inability to make causal conclusions but the data. For example, ANOVA, which is usually used with experimental data, is no stronger than SEM or regression in the presence of non-experimental data (Fabrigar et. al.,2010). However, SEM has advantages when non-experimental data are present. One advantage is the ability to model different competing plausible models. Each model is compared using various statistical tools to find the best model for the data. If one model is better than the other, then the researcher can claim one causal assumption is better than the other. A second advantage is the ability to handle the threat of other alternative variables being left out of the model. When known, multiple regression can control for the effect of these type variables. In SEM however, a measurement error term is used to account for random measurement error and also for systematic error. In fact, SEM often provides a more accurate effect of a hypothesized causal variable controlling for the effects of other causal variables (Bollen, 1989, as cited in

Fabrigar et.al.,2010). Another advantage is SEM's ability to allow for the simultaneous estimation of all the effects in the model. Instead of modeling a series of multiple regression equations, a single model is modeled and ran simultaneously which increases explanatory power in model testing. SEM allows for these simultaneous regression equations by allowing one variable to be both a dependent and independent variable in the same model (Bowen & Gau, 2012). Nusair and Hua (2010) found SEM had a greater number of statistical relationships than ordinary least square analysis (OLS) (as cited in Dwyer, Gill, & Seetaram, 2012). SEM has the ability to handle multiple types of data such as censored, count, non-normal, categorical, and time series (Recker, 2011). Also, SEM allows you to work with latent variables (non-observable variables). When used in measurement model, these latent variables (i.e., factors) are allowed to correlate. The correlation of independent variables in the regression model complicates the interpretation of the regression coefficients (Bowen & Guo, 2012). Lastly, mediation analysis done with regression based analysis leads to inflated estimates of direct effects and attenuation of the mediated effects (Baron & Kenny, 1996, as cited in Fabrigar et. al. 2010). Mediation with SEM allows all types of variables (latent or observed) to be included in the mediation analysis. Iacobucci, Saldanha, and Dang (2007) noted that SEM’s ability to detect mediation structures that exists in the data are an advantage over regression. SEM does have one major drawback, many models are plausible. Therefore, models have to be based on sound theory. Also, SEM requires larger sample size than do regression based models (Kline, 2010).

Structural Equation Modeling (SEM)

Behavior patterns or emotions are unobserved variables that can't be directly measured nor accurately measured with only one questionnaire item. Researchers who want to study the relationship between observed and unobserved variables are left with few options in ordinary statistics. SEM is a broad collection of statistical techniques. One can use it to test the relational structure between a latent construct and observed variables (i.e., factor analysis), the relationship between observed variables (Path Analysis/Regression), or the causal relationship between latent variables (full structural model). All these models (and more) are under one umbrella, structural equation modeling.

One of the earliest forms of SEM was Path analysis. Path analysis is a multiple regression technique that was founded by geneticist Sewell Wright (1918) as a way of testing propositions on non-experimental data (Wright, 1932). By studying the relationship between gestation period and weights of guinea pigs at birth, he utilized correlations and hypothesized regressed relationships between observed variables to produce an acceptable model. Psychologist Charles Spearman (1927) was able to show mathematically that several different intellectual abilities had one common cause (i.e., factor). He was able to use correlations between the observed variables to mathematically discover relationships that existed between the observed variable and their underlying cause (latent variable or factor). Building on the work of Spearman and others, Joreskog (1973), was another prominent contributor to SEM. Joreskog (1973) developed a general model, the LISREL model, for a system of structural equations (Joreskog & Sorbom, 1982).

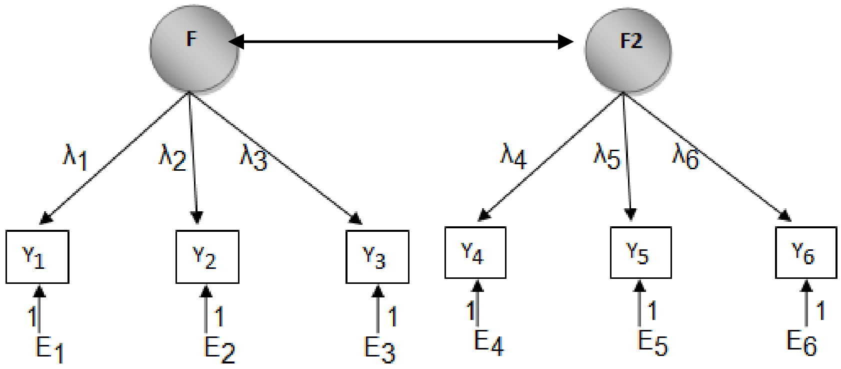
Full structural models are a combination of path models and confirmatory factor analysis. Confirmatory factor analysis models the relationship between the observed variables (indicators) and the factor. If more than one factor is a part of the model, they are allowed to correlate. This is generally called the measurement part of the model. The full structural model is concerned with the structural part of the model (i.e., the causal relationships) between the latent variables. No longer modeled as a correlation, the factors incorporate path analysis via regression like relations with the factor in the model. In SEM, the observed variables are assumed to be multivariate normal; the residual is assumed to be is assumed to be normal with mean zero and variance $\boldsymbol{\theta}$, N(0,$\boldsymbol{\theta}$). The latent factor in a CFA model is assumed to be normal with mean zero and variance $\boldsymbol{\phi}$, N(0,$\boldsymbol{\phi}$). The correlation between the errors is assumed to be zero and the correlation between the errors and factors are assumed to be zero, which is similar to regression where the residuals and independent variables are assumed not to correlate. The model is assumed to be correctly specified, and the relationship between the predictor and dependent variable is assumed to be linear. Note, in SEM, the latent variables and indicators vary between subjects and are assumed independent across subjects. However, in SEM, this is violated since subjects are nested within groups (Skrondel & Zheng, 2007). In path models, variables in models where the regression relationship is modeled on observed variables are assumed to be measured without error.

The goal of each of these techniques is to reproduce the variance-covariance matrix. The closer the model implied matrix is to the original variance-covariance matrix, the better the model. The model implied matrix is composed of the variances and covariances of the observed variables that model a particular relationship. The null hypothesis tests if the population matrix, $\sum$, is equal to the implied matrix,$\sum(\theta)$. Often we have to contend with the sample matrix, so the null equation becomes $H_0 : S = \sum(\theta)$. The sample matrix and the model implied matrix, based on the estimated parameters in the model, are compared using an estimator to minimize differences between the two matrices. A fit function is then is used to quantify which models are best (Bowen & Guo, 2012).

A visual representation of a CFA model is as follows --notice circles represent latent component and squares observed components of the model.

Figure 2.1

CFA model



This SEM model can be represented by six regression like equations for the measurement part of the model. Each observed variable is paired with their respective indicators. The double arrows represent correlation (most times this is represented with a curved arrow). Also note, the observed variable is represented with the letter Y which is not technically accurate. The letter X is used when the observed variable is endogenous, meaning, has an arrow pointing to it. The variable Y is used when at least one arrow is pointed from the variable (i.e., an exogenous variable). The arrow goes from the independent variable to the dependent variable. The symbols used in SEM are Greek or Latin depending on if the variable is exogenous or endogenous, respectively. The system of equations created by the model is as follows,

$$Y_1 = \lambda_1 F + (1)E_1 \tag{13}$$

$$Y_2 = \lambda_2 F + (1)E_2 \tag{14}$$

$$Y_3 = \lambda_3 F + (1)E_3 \tag{15}$$

$$Y_4 = \lambda_4 F_2 + (1)E_4 \tag{16}$$

$$Y_5 = \lambda_5 F_2 + (1)E_5 \tag{17}$$

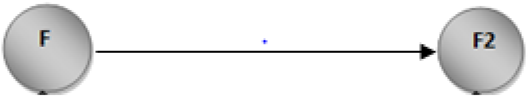
$$Y_6 = \lambda_6 F_2 + (1)E_6 \tag{18}$$

The relationship between a latent factor, F, and one of its' respective indicator is similar to the regression relationship. The latent factor is said to have caused the correlation between the observed variables. The difference between a CFA model and regression model is the lack of intercept and the independent variable is latent. There is no intercept because for a CFA model, the observed variables are entered into the analysis (performed by the program) as deviations from the means. The software centers the variables thereby making the mean of variables zero with no need for an intercept (Bowen & Guo, 2012). Other models, such as those where means are used as inputs, the intercept is modeled.

The loadings are represented by lambda, λ. The loadings represent the strength of relationship between the factor and the indicator. In fact, the product of two standardized path coefficient (lambdas) is equal to the observed variable between the two variables. The correlation Y_1 and Y_2 is calculated as $\lambda_1 * \lambda_2$. For example, if the loading were .50 and .60 respectively then the reproduced correlation would be .30. Also, the standardized loading square represents the explained variance in the variable Y explained by the factor X, which is synonymous to R^2 in regression. So for our example, the first loading explains 25% of the variance in the first indicator but 75% of the variance is unique to the variable and is unexplained by the factor. The measurement error, E, represents unique unexplained variance in the observed variable.

Figure 2.2

Structural part of the model



The structural part of the model no longer has the double arrows but a causal arrow. The structural model is represented as follows,

$$F_2 = \beta F + \zeta \tag{19}$$

The structural part of the model is similar to a regression equal but only with factors as independent and dependent variables. Note, like with anything in SEM there are no hard and fast rules; an observed covariate could be added to the structural equation for an entirely different model. The zeta, ζ , represents the disturbance or factor error.

A good example of a structural equation model is the model for the latent variable social economic status (SES). If our observed variables were income interval, education level, and family's wealth, then all these variables are indicators of a person's social economic status (SES). One of these variables alone would not explain the multi-dimensional construct of SES. The correlation between these variables is explained by SES. If a person has a high level of SES, then they are also most likely to have a high level of income, high level of education, and also a great amount of family wealth.

When dealing with SEM models, there are a number of things one must contend with. For example, when there are more parameters than information. Another issue is scaling. A latent construct or factor has no scale (meters, feet, inches, etc.) in order for the values to make sense of the factor we must attribute a scale to it. There are two scaling options for the latent variable in a CFA model, the reference indicator method and factor standardization. The reference indicator method allows the first loading on each factor to be set to one. The factor standardization method assumes the factor is standardized with a mean of zero and variance one. If the data were continuous, then one of these methods for scaling would be the only requirement for identification of our model. However when the variables are categorical not only do we have to scale the latent factor but we also have to scale the latent response variable, y*. The response variable, y, is categorical but is assumed to be a continuous response variable, y*, divided by thresholds that determine the category. There are two conventions for scaling the latent response variable: marginal parameterization and conditional parameterization (Kamata & Bauer, 2008). The marginal or total parameterization sets the variance of y* to one so now when we estimate the variance of the errors, the equation is $V\left(E_i\right) = 1 - \lambda_i^2 \; Var\left(F\right)$ (Kamata & Bauer, 2008). The second method of scaling the latent response variable assumes the variance of errors is one, $V(y_i^*) = \lambda_i^2 \mathrm{Var}(F) + 1$. This type of scaling is similar to the cumulative probit model and is called the conditional parameterization method since the conditional distribution, (y*|F), is assumed to be standardized (Kamata & Bauer, 2008). The cumulative probit model assumes the conditional distribution of y* follows a normal curve with mean zero and variance one, the values for a standard normal. There is another normative technique used to scale the latent continuous response variable, namely, logit. The logit model normalizes the conditional distribution of y* by assuming it follows a logistic curve and is a logistic random variable with variance of $\pi^2/3$. Combining the two scalings for the latent factor with those of the scaling techniques for the latent variable, four possible combinations of parameterization emerge

(Kamata & Bauer, 2008). For the two possible conditional parameterizations, the conditional reference indicator approach assumes the variance of the measurement errors is one, the mean or threshold of the first observed value on each latent factor is zero, and the factor loading for the first observed variable on each latent factor is set to one. Conditional standardized factor assumes the factor is standard normal and variance of the error as well as the variance of the measurement error is one.

Multi-group SEM

Rarely in statistics do data come from a homogeneous population--most times data come from different groups of people. For example, in most classroom settings we have a mixture of different races and gender. Various aspects of a factor model might hold for one gender or race but not necessarily for another. The purpose of multi-group SEM is to address issues with heterogeneity in the population where the groups are known (e.g. you know that the data consists of 60% Females and 40% Males). Multi-group SEM tests the equality of the model parameters or components over the several groups. In this model, the intercept is included since means of the observed variables, along with the variances and covariance, are submitted for analysis. The measurement part of the model is comprised of factor loadings, measurement error, and observed means/intercepts. The structural model is comprised of factor means, factor variances, and covariances (as well as structural paths). The test for group differences can be tested over the various parts of the SEM model. However, the measurement part is tested before the structure part. More specifically, factor loading, factor covariances, structural paths, and latent factor means are the things commonly tested (Bryne, 2012).

The multi-group equation is given by,

$$\begin{aligned} Y_{ig} &= v_g + \lambda_g F_{ig} + (1)E_{ig} \\ F_{ig} &= \alpha_g + \zeta_{ig} \end{aligned} \tag{20}$$

where α_g is the latent group mean that can vary across groups, vg is the observed group mean that varies across groups, and ζ_{ig} is the disturbance or factor error.

When studying group difference (in SEM), the word invariance is pervasive. Invariance simply means equivalent or equivalence. The steps to test for multigroup invariance was the result of research by Jöreskog (1971) (as cited in Byne, 2008). His recommendation was similar to a global F test. He recommended the performance an omnibus test of the variance/covariance matrices ($\Sigma_1 = \Sigma_2 = \ldots = \Sigma_g$) over g groups to see if there are group differences. This was followed by more restrictive tests to find where those differences exist (Byrne, 2012). However, Bryne (1988a) found cases where the global null hypothesis may be rejected, yet tests for equivalence of measurement and structural invariance still held. (as cited in Bryne, 2012). According to Bryne(2012), the reason for such inconsistencies was no baseline model present for the variance/covariance structure. In order to establish a baseline model, we first test if the factor structure holds across groups then perform subsequent more restrictive tests. Testing for factorial invariance requires testing invariance in the measurement and the structural model. Measurement invariance holds if we have equivalent factor structure, factor loadings, item intercepts, and measurement error variance. Structural invariance is equivalence of the factor means, factor variances, and covariances. Measurement invariance is tested before, and must hold, before structural invariance is tested. If measurement invariance holds, we are saying difference among observed scores is due to scores on the factor not the respective groups. Essentially, if two people have the same score on a factor, measurement invariance holds if they have the same observed scores. If for the same level of the factor two people have different observed scores, we can reason this is a function of them being different and then measurement invariance does not hold.

The first step in testing measurement invariance is testing configural invariance. Configural invariance or pattern invariance asks if the items that define the factor are the same across groups. Since a factor explains the correlation among the variables or indicators, configural invariance asks if the factor explains the same relationship across groups. For example, if someone decided to define self-esteem in women as being measured by degree of body image issues (primarily weight issues), independence, posture, and eye contact, then the goal in testing configural invariance is to see if the same pattern of loading is consistent across groups and the if the same number of indicators (observed variables) are consistent across groups. Specifically, we are looking to establish a baseline model by looking for the same pattern of free and fixed loading hold across groups and if the same number of indicators exist across groups. Free parameters are parameters to be estimated in the model. Fixed parameters are parameters designated to equal a constant are therefore not estimated in the model.

To test for confgural invariance models are fitted for each group separately. Note, that each model does not have to be identical. The structures do not have to be the same. Partial measurement invariance can still be tested when you have different baseline models in the respective groups (Bryne, 2012). Partial measurement invariance occurs when not all measurement parameters are equal in the various groups. Configural invariance is a multigroup test where both established group baseline models are measured as one group over the common indicators. After testing the baseline models for each group via fit tests and also via the modification indexes (MI), I can then proceed to test for other equivalences. MI suggests how much the Chi-square value would be lowered if the stated and/or unstated relationships in the model were removed or added. Weak factorial invariance holds if factor loadings are equal across groups. Strong factorial invariance builds off the requirement for weak factorial invariance, it holds if the loadings and the intercepts are equal across groups. Strict factorial invariance is even more restrictive, it holds if loading, intercepts, and measurement error variances are equal across groups. Comparing the ever increasing levels of invariance is done by using the chi-square diff test.

The question answered by Multi-group SEM: Does the latent factor have the same relationship with the observed variables across groups? For example, are the variables correlated with the self-esteem factor the same across nationalities?

Multi-indicator Multi-independent Causes Model (MIMIC)

MIMIC are another way to compare group differences on the factor (and/or on a particular indicator). This approach usually uses a dummy variable that can take on the value 0 or 1 to represent the presence or absence of a covariate as the differences in factor means or in intercept value. The number of coded variables is the number of groups minus one. This is similar to multiple regression where the number of dummy variables included in the analysis are G-1. It looks like a CFA with an observed covariate pointing to the factor and/or a particular indicator Note the covariate can be continuous or categorical.

The standard model for the MIMIC model is

$$\begin{aligned} Y_{ig} &= v_g + \lambda_g F_{ig} + (1) E_{ig} \\ F_{ig} &= \gamma_g X_{ig} + \zeta_{ig} \end{aligned} \tag{21}$$

ζ_{ig} D is the disturbance term or error for the factor, X_{ig} is for the dummy variable (a dichotomous indicator), γ_g is the regression coefficient representing the strength of relationship between dummy co-variate, and the factor v_g is the intercept.

Multi-group SEM and MIMIC are similar and different at the same time, both have their positives and negatives. MIMIC allows the use of a smaller sample size since you are working with the total covariance matrix. ,In multi-group SEM, you have covariances and variances as well as means for each group. Therefore, each group has to have enough people to run the particular model (dividing the covariance matrix can make this difficult). Also, MIMIC has less parameters to estimate, MIMIC uses the total variance covariance matrix and instead of means adds an additional variable (i.e., the covariate; Brown, 2006). The correlation matrix used by MIMIC involves the covariance between the indicators and the correlation between indicators. The covariate is used as input in the model. A significant beta represents population heterogeneity. Factor means are different at different levels of the covariate. This is similar to testing for factor mean differences for the multi-group SEM model. MIMIC is limited in that it only tests for invariance of intercepts and factor means. The other types of invariances can't be tested using the MIMIC model. Multi-group SEM allows for partial measurement invariance across groups whereas MIMIC does not allow such flexibility (Thompson, Green, 2006).

The question answered by MIMIC: *On average do the covariate variable (race), blacks and whites, differ on the level of the factor (self-esteem)?* This is a question on latent differences for the covariate. However, another question can be asked when the co-variate points to the indicator, DIF (differential item testing). *On average for blacks and whites, do they have different average scores on the indicator (average test scores)?*

Multilevel Structural Equation Modeling

Difference between MIMIC, Multi-group CFA, and Multilevel CFA. Because multilevel CFA models encompass CFA models, multi-group models, and MIMIC models, explaining the various nuances between the various models was necessary. There are inherent limitations of both the MIMIC and SMM models that make multilevel modeling necessary. In general, MIMIC models assume strict invariance across groups while being more parsimonious and requiring less sample size. Multi-group CFA is more flexible, needing only partial measurement invariance (which allows you have invariance on a partial set of indicators), but it requires a larger sample size and can only estimate a small number of between group differences or have a small number of groups (Selig, Card, & Little, 2008). Both MIMIC and SMM are used when there are known heterogeneous populations (unknown heterogeneous populations fall under latent class models). However, like most SEM models, there is an assumption of independence. When data are clustered within groups, those within the group tend to behave alike, which causes incorrect inferences (i.e., bias parameter estimates, inflated chi-squares and negatively biased standard errors; Julian, 2001; Bryne, 2012).

Multilevel CFA or SEM is recommended when you have a large number of groups and when data are clustered (Selig, Card, & Little, 2008). Instead of having a fixed number of groups in the model (like Multi-groups CFA), the groups are randomly chosen from the general population. Because the groups are random, the people are randomly chosen. Therefore the factor means vary across groups and are considered random. Having a random selection of groups allows one to generalize about the inferences that are made (Selig, Card, & Little, 2008). Also, one can think of random as meaning each time you draw you might get a different group.

So, instead of having to make your conclusions limited to the groups in your model, a general conclusion could be made for all such groups. Unlike conventional single level SEM in which independence is assumed over all observations, in multilevel SEM, independence is only assumed over the groups or clusters (Heck & Thomas, 2009). When independence is assumed over people, then knowing something about one person's scores tells me nothing about another person's score. Independence over groups or clusters says something similar, for example, knowing the average score from one school tells me nothing about the average score of another school. Note that the grouping variable has many groups and are called clusters for the multilevel model.

To make a somewhat concise overview, the multilevel model is similar to having a CFA model at the within level and a Multi-group CFA model at the between level. In the standard CFA, there is neither mean structure nor are means included as input, the within level means are not modeled. The between level group means are used as input value and modeled so it is similar to what we think of as the mean structure. Just like there are multiple models in Structural equation models, there are multiple models in multilevel CFA models. The unrestricted ML CFA model is the beginning model used to ascertain if there is between group variability, if the between group should be modeled, and what that model looks like (Heck & Thomas, 2009,). As mentioned before, the unrestricted model looks like a regular CFA model at the within level and mean structure model at the between level.

The fact that ML CFA assumes a weak measurement invariance at the between level for the clusters or groups and with multi-group CFA this must be proven is a major difference between multi-group CFA and ML CFA (Selig, Card, & Little, 2008), The reason for this is because there was no straight forward way to test for measurement invariance at the between level because thirty plus groups or clusters yielded unwieldy results and lead to an increase in type I error (Selig et. al, 2008). Thus multilevel CFA models have only one group at the within level since means are not modeled at that level. The grouping variable at between level is random and not fixed like multi-group CFA. That is with this means model a baseline group is selected and all other groups are compared to the baseline group. Weak measurement invariance is assumed. This is not the only multilevel model that SEM has available. Invariance can be tested across the two levels, at the within level, and at the between level (Selig et. al., 2008).

For the unrestricted ML SEM model, the goal is to confirm that multilevel should be modeled and to decided what the model structure at the between and at the within levels should look like. These models need not look identical at the respective levels. To confirm that multilevel modeling is needed, Muthén's ICC is calculated for each observed variable in the model. Shrout and Fleiss (1979) provided the classical definition of ICC as the ratio of the between group variance to the total variance (as cited in Wang, Xei, & Fisher, 2011).

$$ICC = \frac{{\sigma_b}^2}{{\sigma_b}^2 + {\sigma_w}^2} \tag{22}$$

Muthén defined ICC as the correlation of two individuals in the same group. His definition was based off the classical definition of item reliability. Item reliability is the extent to which the variance of the observed variable is explained by the factor (Wang & Wang, 2012). Reliability is the ratio of explained variable to total variance,

$$Item\,\mathrm{Re}\,liability = \frac{\lambda^2 Var(F)}{Var(Y)} = \frac{\lambda^2 Var(F)}{\lambda^2 Var(F) + Var(E)} \tag{23}$$

If we are looking at completely standardized data, then y and F are N(0,1) and therefore the variance of the factor, var(F) is one and variance of y is one. The reliability ratio then becomes the square of the loadings, λ^2. The loading square is equivalent to R-square or the square multiple correlation, and measures the percent of variance in the observed variable explained by factors for which it loads. Therefore, if the standardized loading is .29 then the factor explains (29%)^2 or 8.41% .of the variance in y (the observed variable; Wang & Wang, 2012). For a given variable y, its' variance can be decomposed into a between and within level (assuming we are talking about a one factor multilevel CFA).

$$Y = Y_b + Y_w \tag{24}$$

$$Y_{ci} = v_{bc} + \lambda_{bc}\eta_{bc} + \varepsilon_{bc} + \lambda_w\eta_{wci} + \varepsilon_{wci} \tag{25}$$

$$Y_b = v_{bc} + \lambda_{bc}\eta_{bc} + \varepsilon_{bc} \tag{26}$$

$$Y_w = \lambda_w\eta_{wci} + \varepsilon_{wci} \tag{27}$$

$$\eta_{bc} = \alpha_b + \zeta_{bc} \tag{28}$$

$$\eta_{wci} = \zeta_{wci} \tag{29}$$

v_{bc} is a vector of intercepts at the between level and varies by cluster (groups are not referred to as clusters, c, for cluster), λ_{bc} is the vector of between level factor loading for y, η_{bc} is the random between level factor that capture the organization/school level effect, ε_{bc} is the between level residual, λ_w is the vector of within or individual level loading, and η_{wci} is the within level factor scores that vary across individual. ε_{wci} is the within individual measurement error which is the error unexplained by the within level factor. ζ_{bc} is for random variance in the between group factor. ζ_{wci} is for random variance in the within group factor, and α_b is the grand mean for the between group factor (mean of means, since we have means at the between level).

Muthén (1991) defines within and between reliability in terms of item reliability. The within reliability and between reliability are represented as

$$\frac{\lambda_w{}^2 Var(F)_{wi}}{\lambda_w{}^2 Var(F)_{wi} + Var(E)_{wi}} = \frac{WF}{WF + WE} \tag{30}$$

$$\frac{\lambda_b{}^2 Var(F)_{bc}}{\lambda_b{}^2 Var(F)_{bc} + Var(E)_{bc}} = \frac{BF}{BF + BE} \tag{31}$$

Using this new nomenclature, he then formally defined ICC for a variable y and individual i and i* in cluster c as,

$$Corr\ \left(y_{ci},\, y_{ci*}\right) = \left(\frac{BF + BE}{\left(BF + BE\right) \ + \ \left(WF + WE\right)}\right) \tag{32}$$

Muthén's ICC describes the correlation of individuals in the same group (Muthén, 1991). If families or schools or groups are highly correlated that means they then have to answer or behave in the same way so their individual scores are not independent, meaning a multilevel model was necessary. Independence between groups is assumed. According to Dyer, Hanges, and Hall (2005), if Muthén's ICC is less than .05 then multilevel modeling should not be used. Julian (2001) stated that even if the ICC is less than .10 the hierarchical structure should not be ignored (as cited in Byrne,2011).

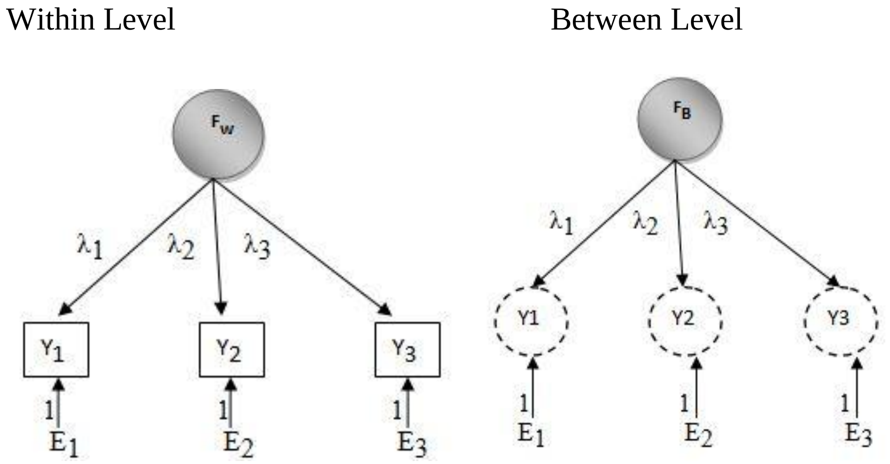
Another way of thinking about Muthén's ICC is that we are testing if the difference between variance/covariance matrix is zero. With multilevel modeling the total variance/covariance matrix is decomposed into the between cluster variance/covariance matrix and the within group variance/covariance matrix. If the between cluster variance/covariance matrix is zero then the within variance/covariance matrix is essentially the same as the total variance/covariance matrix, and it means that there are no between cluster differences ($y_{bc} = \overline{y}$). So for example, if the between cluster variance/covariance matrix is zero and we collected data on BMI as an indicator of self-esteem then this would mean there is no difference on the BMI measure across ethnicities. Suppose we obtain an ICC of .10 that would be interpreted as 10% of the variance in y is at the between level. Muthén's ICC is an estimate of the observed ICC since it is using components of the model (lambda, variance of factor, variance of error) which does not completely replicate or match the variance of y but is an estimate for the given model.

Obtaining the ICC is one of many steps in the establishment of the unrestricted model. Originally Muthén has a four step procedure for fitting multilevel SEM model. Step one fits the total covariance matrix which is the same as doing a regular CFA model to test for general model fit or the approximation of what the model should look like. During this phase you look at the output's model fit parameters like Chi square, such as CFI (comparative fit index) and SRMR(root mean square residual) for model fit. You look to see if the loadings are significant. If the factor loading is not significantly different from zero this signifies that the indicator and the factor are not related. The modification indexes are checked to see if a relationship needs to dropped or considered in the model. The SEM model should be based on theory but sometimes the model does not fit the data so the model might be mis-specified. Modication Indices (MI) gives you suggested ways to improve model fit by freeing model parameters. A MI indicates the decrease in Chi square with 1 df if a particular parameter is freed from a constraint in the model (Wang, 2012) One starts by looking at which MI are very large. The MI indicates the amount the Chi square would drop if we add that particular indicator relationship on that particular factor. If the literature backs up the recommendation then one can make the modification and refit the model. After achieving an acceptable model, the subsequent steps were to estimate the ICC, confirm the within group structure, confirm the between cluster structure, and lastly, the multilevel structure (Dyer et. al., 2005). This process has since been streamlined into a three step process (Bryne, 2011). The new process still involves modeling a CFA model on the total covariance matrix, however, the next step calls for estimating the between and within level simultaneously (originally step five). During this time the structure of the model at the between level and within level are looked at to confirm the number of loading and factors at each level. The model does not have to be the same at each level. One can have two factors at the within level and one at the between level. Once the structure and dimensionality is determined, then the ICC is checked as the final step (Byrne, 2011). The interpretation at both the between and within level model is similar to a CFA model. For a CFA model we are looking to see if the correlation of the variables is explained by the factor or factors. For the within level CFA model, we are looking to see if the correlation between the variables that we see within the groups is explained by the factor. For the between level CFA model, we are looking to see if the correlation on the variables that we see between the groups is example by the factor. For example, using a one dimensional factor model, where we're given general questions on school violence as items, the within level might seek to see if the correlation between the variables within the school is explained by a general bullying factor but the between level would seek to ascertain if the correlation between schools on these (or a subset of these) variables is explained by a general school climate factor.

The question answered by unrestricted Multilevel CFA model: (Within Level) *Does the within level factor(s) explain the correlation of the variables/items within cluster/group?* **(Between Level)** *Does the between level factor (rarely factors) explain the correlation found between clusters?*

Figure 2.3

Multilevel CFA


Within Level

Between Level

The circles at level two represent the absence of raw scores but observed group means given at level two.

A hypothetical data example would be a vector of observed scores on y_{ic}, where we have three clusters and two people per cluster. For this example every two numbers are a cluster, so, 5 and 6 are in cluster one, and 7 and 8 are in cluster two. The second matrix represents the between level vector where the group means vary across clusters. We use this matrix in the calculation of between group variance/covariance matrix.

$$y_{ic} = \begin{pmatrix} 5 \\ 6 \\ 7 \\ 8 \\ 7 \\ 4 \end{pmatrix},\ \bar{y}_{ic} \begin{pmatrix} 5.5 \\ 7.5 \\ 5.5 \end{pmatrix} \tag{33}$$

After testing the unrestricted model, the next model that can be tested is the model that measures measurement invariance across levels. In this model, the between level is treated as one group and the within level is treated as one group (Bryne, 2001; Heck & Thomas, 2009). By testing this restricted model we can find out more about the latent variance at the between level vs. the within level, latent ICC. The previously discussed model helped us ascertain the Muthén's estimation of the observed ICC. In general, for a multigroup CFA model, we try to establish measurement invariance to ascertain if a construct is measured the same way across groups. Similarly, we are trying to establish if the latent factor is measured the same way across levels. Since each level is considered a group, we are doing a two group test for invariance. If measurement invariance fails, then the meaning of the construct is not the same across group (i.e., the factor has different meaning across levels). To test for measurement invariance across levels we must establish a common factor that has a between and within level (Heck & Thomas, 2009; Selig, Card, & Little, 2006). According to Mehta and Neale (2005), "Invariant factor loading makes the common variance attributed to the latent factor directly comparable across levels"(as cited in Heck & Thomas, 2009, p. 124).

$$\eta_{ic} = \alpha + \eta_{bc} + \eta_{wci} \tag{34}$$

η_{ic} represents our common factor that has a grand mean, α , and a between cluster and within cluster factor. Just like each observed indicator, Y, was split into a between and within level, the factor is now split into a between and within level.

The common factor model is one where the between and within factor represent one factor, η_{ic} . To develop a common factor model loading, λ , are assumed to be equal across levels. Note, if there are problems with estimation, the between level error variance can be constrained to zero for some of the problem indicators. This is okay because the between level error is usually small (Heck & Thomas, 2009, p. 119). "In the between-group level model, residual variances are typically very small, which reflects high reliability" (Heck & Thomas, 2009, p.119). Thomas and Heck (2009) did say that this should be based on theory and should not be changed for the sake of changing models, "We emphasize that model modifications should be made sparingly and with regard to theory and statistical power" (Thomas & Heck, 2009, p. 126).

The original multilevel model looks like $Y_{ci} = v_{bc} + \lambda_{bc}\eta_{bc} + \varepsilon_{bc} + \lambda_{wci}\eta_{wci} + \varepsilon_{wci}$. After restricting the loadings to be equal at the between and within level you can observer the grouping of the factors,

$$Y_{ci} = v_{bc} + \lambda\eta_{bc} + \varepsilon_{bc} + \lambda\eta_{wci} + \varepsilon_{wci} \tag{35}$$

$$Y_{ci} = v_{bc} + \lambda(\eta_{bc} + \eta_{wci}) + \varepsilon_{bc} + \varepsilon_{wci} \tag{36}$$

The key advantage to modeling the common factor model is that we can now ascertain the between cluster variance, η_{bc}, the true or latent ICC.

$$LatentICC = \frac{BF}{BF + WF} \tag{37}$$

Looking at the restricted model, Thomas and Heck (2009, p.123-124), set the loading to equal at the within and between levels, and only set one between level residual error to zero (to help the model converge). After doing so, he found the model converged and the between factor had 20.8% of the variance. In other words, 20.8% of the variance in the common factor is at the between level.

The question answered by the restricted Multilevel CFA model: *Is the relationship with the factor the same at the within and between level?*

There are many more models that can be added, such as a multilevel MIMIC model that can add a co-variate to the within or between level. These are still new, forthcoming research areas.

Thus far, we have talked about the latent variable and the observed variable being broken into the within and between group estimates. Variance/co-variance matrix is an integral part of

SEM so focusing on this aspect is critical. The total population covariance matrix is written as a function of the population between cluster covariance matrix and the within cluster covariance matrix. The within correlation matrix represents covariation at the individual level, individual difference while controlling for cluster. The between correlation matrix represents covariation at the cluster level (i.e. differences across cultures on the factor).

$$\sum_T = \sum_W + \sum_B \tag{38}$$

If there is no between level variance/covariance, then the within level covariance matrix would equal the total covariance matrix and we would only need to do single level CFA. Muthén (1989) showed that the sample pooled within group covariance matrix, S_{pw}, is an unbiased estimate of the population within group covariance matrix. We can estimate the within group by constructing this matrix: $S_{pw} = (N-C)^{-1}\sum_{c=1}^{C}\sum_{i=1}^{n_c}(y_{ci}-\overline{y})(y_{ci}-\overline{y})^T$.

He showed that the scaled sample between group covariance matrix, S_b, is not an estimate of the population between group matrix, $\sum_B$, but a consistent and unbiased estimator of $\sum_W + c^* \sum_B$.

$$S_b = (C-1)^{-1}\sum_{c=1}^{C} n_c(\overline{y}_c - \overline{y})(\overline{y}_c - \overline{y})^T \tag{39}$$

where c* reflects common group size, if balanced, common group size. C is the total number of clusters. For unbalanced data and large number of groups, c * is close to the mean of the cluster sizes.

$$c^* = \left[N^2 - \sum_{c=1}^{C} N_c{}^2\right]\left[N(C-1)\right]^{-1} \tag{40}$$

The maximum likelihood (ML) estimate of $\sum_W$ is S_{pw}, while the ML estimate of $\sum_B$ is $c^{-1}(S_B - S_{PW})$. Muthén (1990) used a full information maximum likelihood (FIML) function to fit the parameters for an ML CFA model. Using $\sum_w$ and $\sum_b$, to represent the model implied or estimated by the between and within cluster covariance matrix. The fit function for balanced design model is

$$F = C\left\{In\left|\Sigma_w + c*\Sigma_b\right| + trace\left[(\Sigma_w + c*\Sigma_b)^{-1}S_b\right] - In\left|S_b\right| - p\right\} + (N-G)\left\{In\left|\Sigma_w\right| + trace\left[(\Sigma_w S_{pw})^{-1}\right] - In\left|S_{pw}\right| - p\right\} \qquad (41)$$

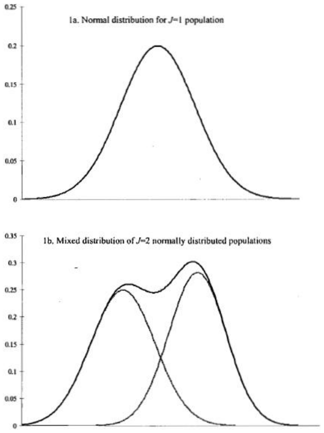
N is the total number of observations, C the total number of groups and p is the number of variables or indicators in the model.

Mixture Models

When a set of data comes from a mixture of multiple populations that have differing univariate distributions (means and/or variances), this is referred to as mixture modeling. As you see from the figure below two normal populations were mixed to obtain a new bimodal distribution (Gagne, 2006).

Figure 2.4

Mixture of two Normals



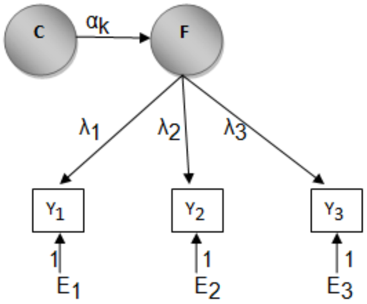
With the models we have seen thus far, multi-group and MIMIC, the number of groups was known. This is not always the case in SEM. In SEM, when the number of groups that make up this mixture is unknown, we use either a parametric mixture model (factor mixture model) or non-parametric mixture model (latent class model) to find the number of groups causing the heterogeneity in the data. The factor mixture model assumes the factor is continuous while the latent class model (LCM) assumes the factor is categorical. Population heterogeneity represents the presence of two or more latent or unobserved groups in the population that have different distributions (Lubke & Muthén, 2005). Muthén (2002) was able to create data using a CFA model that was a mixture of two distribution on the continuous factor by manipulating the proportion of people in each group and by creating a different mean and variance (or distribution) for each group. In essence, what he did was reverse factor mixture modeling. His objective was not to run a factor mixture model to find the number of groups causing the heterogeneity, but to create the heterogeneity while using only two groups with different distributions.

In the above picture, we could see that we had bimodal distribution but the number of groups is unknown. Therefore if we were just given the bimodal distribution that we suspect is formed by k normal classes, we could use log likehood functions to estimate the number of classes, as well as the parameters for the each of the normal populations, namely, mean and variance.

Log likelihood estimation is a tool used to estimate the parameter values for the population when you are given a sample. For a normal distribution, the parameters are the mean and variance. So for a given set of data points or sample, I can use the probability density function for the distribution with various estimates of the unknown population parameters. The best estimation of the population parameters (i.e., mean and variance) are the ones that yield the highest probability value or rather maximize the probability of seeing the sample drawn.

Figure 2.5

Factor Mixture Model



Lubke and Muthén(2005) describes a one factor mixture model as follows:

$$Y_{ik} = \lambda_k \eta_{ik} + \varepsilon_{ik} \tag{42}$$

$$\eta_{ik} = \gamma_k C_i + \zeta_{ik} \tag{43}$$

$$C_{ik} = \begin{cases} 1, if\ i \in class\, k \\ 0, if\ i \notin class\, k \end{cases}$$

C is an indicator or dummy variable that is 1 when you are in class k and 0 when you are not in class k and there are k=1,.....,K classes. Lambda, λ_k, is the factor loadings that are assumed invariant across classes so differences on the mean factor scores can be explored. γ_k is a vector of factor means.

The normal probability density function is

$$f\left(x_i\right) = L_i = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)e^{\frac{(-.5)(x_i-\mu)^2}{\sigma^2}} \tag{44}$$

where the indexed x variable represents a single sample observation on the variable X, μ is the mean of the X in the population and is the population variance. The density function is the probability of observing a value of x_i from a normal distribution with those parameters chosen. For a sample size of N, assuming independent observations, the joint likelihood is the likelihood for the entire sample and is calculated as a product of all the likelihood functions, $\prod_{i=1}^{N} L_i$ (Gagne, 2006). Using the fact that the log of a product is equal to the sum of the log, $\prod_{i=1}^{N} L_i = \sum_{i=1}^{N} In(L_i) = \Psi$, the derivative of the new form likelihood function is then taken and set to zero in the final step of finding the parameter values.

For the mixture model, each normal distribution has a likelihood function. If we knew a priori that that the observed distribution was from the mixture of two normal populations, then a new joint likelihood function would emerge that is a weighted sum of the two likelihood functions

$$L_i = \varphi L_{i1} + (1-\varphi)L_{i2} \tag{45}$$

where φ is the mixing proportion of the number of people thought to be in the respective populations. Once the new likelihood function is estimated, not only is the parameters estimated for each of the respective likelihood functions but also the mixing proportions (Gagne, 2006). Information criteria, such as the AIC (Akaike information criterion) and BIC (Bayesian information criterion, that measure fit, are used to determine the number of normal populations present. The model with smallest AIC or BIC value is chosen and therefore indicates the correct number of normal populations present.

$$\Psi = \sum_{i=1}^{N} In(L_i) = \sum_{i=1}^{N} In(\varphi L_{i1} + (1-\varphi)L_{i2}) \tag{46}$$

$$\begin{aligned} AIC &= -2\Psi + 2q \\ BIC &= -2\Psi + In(N)q \end{aligned} \tag{47}$$

So for a mixed population, Ψ is maximized, and q represents the number of parameters to be estimated (if the distribution is made up of two populations then you have to estimate two means, two variances, and one proportion; the other proportion is the complement which therefore sums to five parameters; Gagne, 2006, p. 204).

Modeling mixture models in a common factor model (CFM) is different because we have to change the parameters to fit the specifics of our model.

$$\begin{aligned} X_i &= \tau + \Lambda\xi_i + \delta_i \\ \mu_x &= E(x) = \tau + \Lambda\kappa \\ Var(X) &= \Sigma = \Lambda\Phi\Lambda' + \Theta \end{aligned} \tag{48}$$

The standard equation for the CFM, for a particular variable X, includes the vector of the grand mean of X, τ ; the matrix for the factor loading, Λ ; the vector of the latent factor values which vary over individuals, ξ ; and also the vector of the measurement error value which also varies over individuals, δ . The first moment is the mean, given by μ_x , and the second moment is the variance.

For a single population, the maximum likelihood function, is assuming multivariate normality, is

$$\prod_{i=1}^{N}(2\pi)^{-p/2}\left|\Sigma\right|^{-1/2} e^{\frac{(-.5)(x_i-\mu_x)'\Sigma_j^{-1}(x_i-\mu_x)}{\sigma^2}} \tag{49}$$

and for a multisample distribution, the likelihood function becomes,

$$\prod_{j=1}^{J}\prod_{i=1}^{N}(2\pi)^{-p/2}\left|\Sigma_j\right|^{-1/2} e^{\frac{(-.5)(x_i-\mu_{xj})'\Sigma_j^{-1}(x_i-\mu_{xj})}{\sigma^2}} \tag{50}$$

where j is the number of multivariate normal distributions, and p is the number of observed variables for a indicator variable X. The model implied variance matrix is given by $\sum_j$. For a sample common factor model, the maximum likelihood fit function is given by,

$$F = \left[In\left|\Sigma\right| + tr(S\Sigma^{-1}) - In\left|S\right| - p \right] + (m-\mu_x)'\Sigma^{-1}(m-\mu_x)' \tag{51}$$

where m is the observed means and $\hat{\mu}_x$ is the model implied mean vector

For a multisample model, the maximum likelihood fit function becomes

$$G = \Sigma_j \left({n_j}/{N} \right) \left\{ \left[In\left|\Sigma_j\right| + tr(S_j\Sigma_j^{-1}) - In\left|S_j\right| - p \right] + (m_j-\mu_{xj})'\Sigma_j^{-1}(m_j-\mu_{xj})' \right\} \tag{52}$$

Note ${n_j}/{N}$ is the proportion; so the multisample is the sum of the proportional likelihood over j populations. Also, maximizing the log likelihood function is equivalent to minimizing the fit function in SEM.

For my study, skewness and kurtosis play more of a role when trying to understand a mixture distribution. For a mixed distribution model, Mplus (Tech 12, 2004), theoretical skew and kurtosis is given by

$$Skew(Y) = \frac{E(y^3) - 3E(y^2)E(y) + 2E(y)^3}{Var(y)^{1.5}} \tag{53}$$

$$Kurt(Y) = \frac{E(y^4) - 4E(Y^3)E(y) - 3E(y^2)^2 + 12E(y^2)E(y)^2 - 6E(y)^4}{Var(y)^2} \tag{54}$$

where the kurtosis is centered at zero. The sample skew and kurtosis is given by

$$\begin{aligned} SSkew(Y) &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \overline{Y}}{\sqrt{S(Y)}}\right)^3 \\ SKurt(Y) &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \overline{Y}}{\sqrt{S(Y)}}\right)^4 - 3 \end{aligned} \tag{55}$$

SEM estimators

The goal of SEM is to obtain parameter estimates based on minimizing the distance between the sample covariance matrix and model covariance matrix (the model matrix is the one you postulate based on theory). The fit function, F, represents the distance between the sample co-variance matrix and matrix based on your model. Estimation methods minimize the fit function.

There are numerous ways to write the fit function, however, each representation involves W, a weight matrix, which is used to “correct for bias in the standard errors and fit statistics,” S (the sample covariance matrix based on the observed data), and the covariance matrix based on your model (Schumacker & Lomax, 2004, p. 31). There are various estimation methods that are differentiated by how they define the weight matrix. Maximum likelihood (ML) is the best known estimation method employed; it uses the inverse of the model matrix, ∑-1, as its weight matrix. The maximum likelihood estimation method is an iterative approach to finding parameter estimates. The parameter estimates are estimated by maximizing the probability of observing the sample under an assumed normal distribution. One of the disadvantages in using maximum likelihood to estimate your parameters is that it assumes multivariate normality, no missing values, a correctly specified model, and independent observations (Kline, 2005). Studies have shown Browne’s asymptotically distribution free (ADF) estimation methods correct for non-normality by using methods of moments to accommodate kurtosis (measure of flatness) in the weight matrix. The Satorra-Bentler (SB) chi-square statistic and SB Robust Standard Errors are alternative ways of adjusting for non-normal continuous indicators by adjusting the standard errors and chi-square values. ADF estimator works best for very large non-normal data. All the above estimators and/or adjusters assume continuous data. In real life, data are rarely normal, rarely independent, and rarely continuous. When the data are non-continuous, categorical Weighted Least Squares (WLS) is often recommended as an estimator (Brown, 2006). WLS is class of estimation that takes on various implementations depending on the software. For example, SEM software Mplus, uses WLSM and WLSMV (which are a mean and variance adjusted WLS estimators).

Like many general linear models, SEM assumes normality and continuous observed data; however, Micceri (1989) found that non-normality is rarely the case in psychological and educational data. Unlike many general linear models, SEM allows you to choose your estimators. Estimators are generally evaluated on three criteria: bias, consistency, and efficiency. An unbiased estimator correctly estimates the population parameter without error; it does not underestimate or overestimate (Finney & DiStefano, 2006). A consistent estimator is an estimator that converges to the population values as the sample size increase; it is asymptotically equivalent. An efficient estimator has the smallest variance; “variability of the parameter estimate is at minimum in large samples” (Finney & DiStefano, 2006, p.271). In statistical theory, these properties make up the criteria we use to choose the best estimator for a population value. The estimator with these properties is said to be BUE (best unbiased estimator) or BLUE (best linear unbiased estimator).

Kurtosis and skewness are measures of normality. A normal distribution or bell shape curve has its highest peak at the mean and it is symmetric about that point. Kurtosis is a measure of peakedness or flatness of the distribution relative to the normal curve whose kurtosis is zero. Skewness is a measure of symmetry relative to the normal curve whose skewness is zero. According to Finney and DiStefano (2006), there is no guideline for what is acceptable kurtosis or skewness, but normal theory based estimators such as ML start to break down when univariate kurtosis or univariate skewness is between two and seven. This break down can affect parameter estimates, standard errors, fit indices, and chi-square statistics. The chi-square statistic becomes inflated under conditions of moderate non-normality and increases as non-normality increases. Parameter estimates are not affected, and standard errors are underestimated thereby inflating the test statistic (Finney & DiStefano, 2006).

According to Bollen (1989), coarse categorization of continuous variables is common in social science. For example researchers, for convenience, may shorten a variable such as years of education into three or four categories such as high-school, college, and graduate-school. The loss of variability can affect the precision of an estimate under certain distributional conditions. In general, if the number of categories is at least five for normal data then we can treat the categorical variable as continuous. Green et al.(1997) noted that as the number of categorization increased for normally distributed distributions, the chi-square values approached those of continuous data (Green et al., 1997). As stated before, normality is unlikely in social science. When non-normality is coupled with non-continuous data, we have more challenges. Under these two conditions, the ML based chi-square value is inflated. The non-normed fit indices (such as

NNFI, CFI, etc.) are underestimated and underestimation of parameter estimates and standard errors become more apparent as univariate kurtosis and skewness increase (Babakus et al.,1987, as cited in Finney & DiStefano, 2006). ADF estimators, robust WLS methods, SB scaled chi-square and standard errors are some of the techniques we use to correct the bias and inflation caused by categorical data. Which strategy you use to correct for non-normality or non-continuous data depends on the limitation of the software program. Mplus, a popular statistical software that is often used in modeling SEM models; Mplus implements SB adjusted WLS by using WLSMV and WLSM (Finney & DiStefano, 2006).

Model misspecification can also affect estimates depending on which estimator is used. Bandalos (2011) studied the performance of estimators under model misspecification, non-normality, various sample sizes, and course categorization. She found that under the optimal conditions of larger sample sizes, more categories, and more normally distributed data convergence was not a problem, but those models with large numbers of parameters, small numbers of categories, and low sample size had convergence issues. WLSMV parameter estimates and standard errors were least affected by model misspecification. Maximum likelihood SB adjusted Mplus estimator (MLMV) and WLSMV underestimated standard error as non-normality increased but performed better than their non SB adjusted counterparts.

Most of the studies have estimated the effect of non-normality or categorical data in the context of single level SEM data. Multilevel SEM is advocated when the independence assumption is violated. Although heterogeneity is thought be controlled by modeling the data as a multilevel model, the effect of non-normality, sample size, and continuity still affect the estimates, even in the case of multilevel data. The question is how? In multilevel data, the effect of sample size and un-balanced group sizes at the higher levels can affect conversion, the parameter estimates, and standard errors of between level estimates. Hox et al. (2010) studied multilevel SEM under the assumption of multivariate normality and no un-modeled heterogeneity by using various estimators (such as robust ML-MLR and WLSMV) under various conditions such as sample size, various ICC conditions, and balanced and unbalanced group size. They found that at least 50 groups were needed for ML, WLSM, and WLSMV; 200 groups was needed for robust ML-MLR; and cluster size had no effect on the accuracy of statistical tests. Meuleman et al. (2009) noted that in many international surveys of countries, many smaller countries are not included in the survey. At most 30 countries are included, so studying multilevel SEM would prove to be difficult even if the nesting of the data call for its use. In their research, they found that the researcher should consider not just the group size at the upper levels in MSEM but also the expected effect sizes and complexity of the model. If the researcher wants to detect small effects, they recommend at least 100 groups. If the researcher has a simple structural model at the between level (i.e., very few parameters to estimate), then 40 groups can be sufficient. In general, when the number of groups is low, say 20, the factor loadings and error variances are underestimated, and the structural effect of the between level is overestimated (Meuleman et al., 2009). Both of these studies assumed normality. According to Hox and Maas (2010) and Preacher (2011), no one has studied the effect of non-normality on MSEM. Since non-normality is known to affect the parameters and standard errors of single level SEM analysis and most of the data for behavior research do not follow a univariate or multivariate normal distribution (Micceri, 1989), this is very timely and much needed research.

CHAPTER THREE

STUDIES TWO AND THREE METHODS

Multilevel SEM is a great tool that allows one to separate the model into between and within level components and control for clustering. Despite the benefits, knowing the right sample size needed to run a multilevel SEM model is critical. Hox et al. (2010) found that the number of clusters needed for multilevel structural equation modeling (MSEM) depends on the estimation method used. They also noted that no one has studied the effects of violating multivariate normality within the framework of sample size requirement for MSEM. This study seeks to become the first study to understand and shed light on this much needed research area of multilevel SEM; and this section gives an overview of how the study was designed and why.

The studies are two simulation studies that seeks to ascertain sample size requirement for various estimators in which the data are non-normal continuous and non-normal categorical, studies two and three respectively. In order to begin studies two and three, a decision on the sample size (i.e., number of clusters or groups), estimation method, average group size/cluster size, and factor structure are just some of the few decisions that had to be made, including the amount of sparseness in the data for non-normal categorical data. Study two seeks to answer three research questions: does sample size requirement for non-normal continuous data depend on estimation method; Is the sample size requirement greater for normal or non-normal continuous data for the respective estimation method; and does the presence or absence of unbalanced clusters affect the sample size requirement for non-normal continuous data. Study three seeks to answer four research questions: does sample size requirement for categorical independent variable data depend on estimation method; Is the sample size requirement the same or different compared to the normal multilevel data for the respective estimation method; does the presence or absence of unbalanced clusters affect the sample size requirement for categorical data; and does the presence of sparse tables affect the sample size requirement?

Using Mplus 6.1 for Linux, Monte Carlo (MC) simulations for these two studies were created by using a Unix script to run multiple conditions at one time. With Monte Carlo studies, samples are repeatedly drawn based on the user defined population values for the model, then the calculation of the parameter estimates and standard errors are based on the average across those samples (Muthén & Muthén, 2002). These averages can be used to determine precision (Muthén & Muthén, 2002). We look at precision in term bias and coverage. Bias describes how far the average statistic is from the population parameter by looking to see if the estimate consistently overestimates or underestimates the parameter. Coverage is the percentage of replications/samples for which a 95% confidence interval covers the population parameter. Muthén and Muthén (2002) outlined three criteria for determining the correct sample size: 1) the parameter and standard error bias should not exceed 10% for any parameter in the model; 2) when the parameter estimate is the focus of power analysis, the standard error bias should not exceed 5%; 3) the coverage should be between .91 and .98%. The coverage describes how accurate the confidence intervals are; we want at least 91% of the 95% confidence intervals to contain the population parameter value. Using these criteria, we can select the best sample size (Muthén & Muthén, 2002).

In a power analysis, we are testing in the null hypothesis if the parameters are significantly different from zero (e.g. if a beta parameter, a path, or latent correlation are significantly different from zero). In general, power is the probability of rejecting the null when the null is indeed false. In a Monte Carlo study, power is the proportion of these replications for which the null hypothesis is rejected (where H_0: β=0). When the population values are different from zero, power is the probability of reject the null hypothesis when it is false. The percentage of significant coefficients is the proportion of replications in which the parameter is significantly different from zero at a .05 alpha level; it represents power when the population value is different from zero. When the parameter has a population value equal to zero, then power is the probability of rejecting the null when it is true. Therefore, the significant coefficient is an estimate of type I error (i.e., probability of rejecting the null when the null is true). Since we are not doing a power analysis, we will look at the first and third criteria as my primary way of assessing sample size. Note, there are other recommended ways of assessing sample size that require different criteria. Later in this section some of those methods will be discussed.

Bias. Parameter bias and standard error bias of the parameter are calculated in the same way. They are calculated by subtracting the population parameter value from the average parameter value over all replications, dividing this difference by the population value, and then multiplying by 100. The formula is as follows:

$$Bias = \frac{\bar{\theta} - \theta}{\theta} \tag{56}$$

where $\bar{\theta}$ equals the mean of the parameter estimates across all the replications, and θ is the true value of the parameter in the population model. This number was multiplied by 100 to get the percentage bias.

Sample size (number of clusters). In standard SEM, the sample size requirement depends on a number of factors including the type of variable, the strength of relationship between the variables, the amount of missing data, the reliability of the variables, and the size of the model. These sample size recommendations are determined during Monte Carlo studies which look at accuracy, precision, and power to determine sample sizes. Precision, or percentage error, is determined by holding the parameter and standard error bias, and seeing if it is below a certain criteria. The precision is assessed by looking at the spread or confidence interval. We want to have a certain percentage of the 95% confidence intervals capture the parameter estimate. The assessment for power of certain parameters determine sample size when power is .80 or better (Muthén ,2002; Raykov, 2006; Brown,2006). Despite numerous possible determinants dictating the best sample size, Raykov (2006) noted that a general rule of thumb is ten times the number of free model parameters because your sample size for traditional SEM is often given. He noted that no rule of thumb can be applied to all situations.

There have been only a few studies that dealt with sample size requirements for multilevel SEM. Hox & Maas (2001) found that the number of groups was more important than the number of people in the group. The estimation method he used was the full maximum likelihood estimation method (FIML), which is the same as MLR today. Hox et al. (2001) used a variety of sample sizes (i.e., 50, 100, and 200), two factors with three indicators each at the within level, and one factor with six indicators at the between level to amass a total of 31 parameter estimates (13 within level and 18 between level). They found that the number of clusters primarily depends on the interclass correlation (ICC) and number of groups. The within part of the model posed no problem, but the imbalanced data posed a problem with model fit. The recommendation from the study was to have a sample size of 100 (i.e., a minimum of 100 clusters) for balanced and unbalanced data with a low ICC: " Given our result, we caution against using multilevel SEM when the number of groups is smaller than 100, especially if the ICC turns out to be low, that is, under .25" (Hox and Mass, 2001, p. 171). Also, if the number of groups is limited by nature, they recommend increasing group size and avoiding extreme, unbalanced data. These recommendations were refuted by both Hox, Maas, and Brinkhuis (2010); and Mueleman and Billiet (2009). Hox et al. (2010) found that the ICC was not an important feature for determining sample size "contrary to the result in Hox and Maas(2001) who found that lower ICCs lead to convergence problems" (p. 166).

Mueleman and Billiet (2009), while studying normal multilevel structural SEM model, also noted that obtaining a sample size (or number of clusters) of 50 is quite impossible, stating that countries on surveys rarely exceed 20 (mostly due to budgeting constraints). Using several variations of their model that had one observed variable regressed on a factor with four indicators, they varied the effect size, model complexity (by constraining some free parameters) , ICC (.08 to .50), sample size or number of clusters (20-100), and average cluster size of 1755 (varied from 1100 to 2800). From this, they concluded that, aside from the impossibility of obtaining a sample size of 50, required group sizes depends on the specific interest of the researcher, expected effect sizes, and complexity of the model. When the number of clusters was 20, there was a very high number of inadmissible conditions; the factor loadings and error variance were underestimated while the structural effect was overestimated. If the between level was relatively simple (meaning a small number of indicators, a max of one structural effect, and no interactions), then a sample size of 40 is sufficient. To detect a large structural effect (>.50) a sample size of at least 60 is required, and more than 100 is required for small effects. These are general guidelines since his study could not cover all SEM conditions.

Hox et al. (2010) reaffirmed that the within-group component for all the models were accurate. The most important factor was the between level sample size. However, imbalance did affect the fit indices, and a minimum sample size of 50 for non-robust estimation methods was still recommended. All of these studies assumed multivariate normality. One of the few studies that researched sample size and multilevel non-normality was Moineddin (2007), who studied multilevel logistic regression models. He found an interaction between prevalence of the outcome, sample size, and group size. His recommendation was a minimum group size of 50 with at least 50 groups to produces valid estimates for multilevel logistic regression models. However, for low prevalent events or events with low probability of occurring, a minimum of 100 groups and 50 individuals per groups were recommended. In all of Hox's studies on multilevel SEM, he consistently used three sample sizes of 50, 100, and 200 for the number of clusters. I intend to follow similar pattern but I will also include a smaller sample size of 30. My chosen sample sizes or numbers of clusters are 30, 50, and 100.

Group size. Hox et al. (2010) after studying the effect of group size, within the context of MSEM, found that group size was not a factor. In this study they equally divided the groups proportionally. Half of the groups would have a small number of people and the other half would have a much larger number of people. The ratio of the number of people in the larger to smaller group was held to around three (e.g., 25 groups of size 3 and 25 of size 7, the average group size was 5 and 7/3 ratio was roughly 3). Their research found that group size only had an effect when the estimation method was the Muthén’s psuedobalanced estimation method. In contrast, Moineddin’s (2007) study about multilevel logistic regression found that group size or cluster size was an important factor, recommending a minimum group size of 50 for events with a low probability of occurring in the population (.10 was the lowest prevalence rate considered).

Hox et al. (2010) average group sizes varied from 5 to 25. Since group sizes might be a factor in non-multivariate normal data, I believe it is critical that my study has cluster sizes beyond what has been studied previously. Three group sizes n=10, 26, and 50 will be included in this study (even numbers were adopted for ease of coding). The final sample sizes, N= nG, will go from 30(10) to 100(50) or 300 to 5000, where n is the number of clusters or sample size and G is the cluster or group size.

Balanced-Unbalanced. My study will consider unbalanced cluster size as part of the independent variables, as well as, balanced cluster sizes. For my balanced clusters, the group sizes were 10, 26, and 50; and the sample size (number of clusters) was 30, 50, and 100. For my unbalanced clusters, I am using the same sample sizes but the group sizes become average group sizes. There will be three average group sizes (i.e., 10, 26, and 50). For the average group size of ten, 50% of the clusters will have size seven, 50% will have size 13, and each will have a 1.85 multiplicative difference. For the average sample size of 26, half of the clusters will have size 18 and the other half size 34, a 1.88 multiplicative difference. For the average sample size of size 50, half of the clusters will be of size 35, the other half of size 65, and each will have a a 1.85 multiplicative difference. The difference was chosen to be roughly similar just in case the ratio of state of unbalance was another uncontrolled variable.

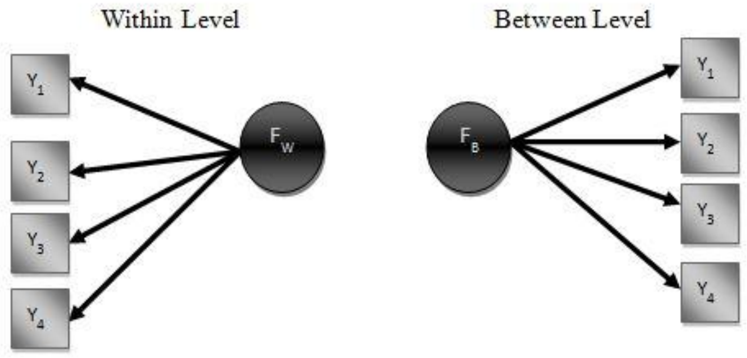
Replications. Before data are generated, the choice of seed and replication are two important decisions that have to be made (Bandalos, 2011). Harwell (1996) noted that the number of replications depends on the purpose of the study. Studying the effect of sample size on parameter estimates versus standard errors might need different replications because more sampling variance would be needed to study the behavior of standard errors (as cited in Bandalos, 2011). For standard SEM, Bandalos (2011) recommends a minimum of 500 reps per cell, deeming 500 reps fairly large for SEM studies and acceptable to obtain stable estimates for standard errors. Meuleman's (2009) study of multilevel SEM models included 25 conditions (5 groups sizes and 5 effect sizes) with 10,000 replications. Hox et al. (2010) had 90 conditions (3 sample sizes, 3 group sizes, 5 estimation methods, and 2 ICC values) with 1000 replications assuming normal distributed latent variables and multivariate normal observed data. Muthén's (2002) study of sample size and power for non-normal and normal indicators for CFA models used 10,000 replications for stable parameter estimates. Moineddin (2007) generated 1000 reps for his study on multilevel logistic regression. To ensure stability of the parameter estimates, I used 1000 reps for both studies two and three.

Seed. A seed is a number that serves as a starting point for random draws used to create the sample generated. A seed can be static or dynamic. A dynamic random seed depends on the computer's internal clock, whereas a static seed is user supplied (Bandalos & Leite, 2011). The advantage of using a static seed is that it is constant, and it serves as a way of controlling for the randomness of the random number generator (Paxton et al., 2001). Bandalos & Leite (2011) do not recommend using the same seed because doing so introduces dependencies; therefore, a repeated measures or some dependent sample design might be needed to test for these dependencies. I used a different seed for each replication for each condition in study two. However, the same seeds were replicated in study three. Using a generic random number generator table, I systematically chose every other number to be the seed within each section of the table.

Factor structure and ICC. Meuleman and Billiet (2009) also studied the conditions of model complexity on sample size requirement. In his multilevel SEM study that included four model sizes or model parameters (namely, sizes 7, 14, 24, 34) crossed with the five sample sizes (creating twenty conditions), they found model complexity was an important factor in determining estimation accuracy, and that reducing model complexity to 7 substantially reduced bias in the parameter estimates. A sample size of 60 was required when there were 7 to 14 between level parameters. To illustrate how he counted model size, the 14 between level parameters included four indicators, one covariate, and one factor. Therefore, four factor loadings, four error variances, four intercepts for the factors, one structural effect, and the mean of the between level independent variable had to be estimated. Hox et al. (2010) used a multilevel CFA model, which did not contain any structural effects. His between level structure consisted of one factor and six variables, and the within level structure consisted of two factors with three variables on each factor (a total of 18 between level free parameters). In multilevel SEM or CFA, the consensus is that a simple between level structure is recommended (Bryne, 2012). In standard SEM, Bandalos (2011) noted that CFA models resulted in less standard error bias than full latent models due to fewer variables, and this effect was attributed to model size not type. She recommended that, for generalization purposes, it might be better to have more than one type of model but when there is little research in an area, as there is in the area of multilevel SEM, it is better to choose one representative model and vary the independent variables (number of groups, group size, ICC, etc.). My model for study two has one factor with four indicators at both between and within-levels, making it somewhat similar to Hox et al. (2010). Study two has the same structure at the between and within levels; there are four indicator variables on one factor at both levels. I avoided model complexity issues by having a maximum of four observed variables, and having one factor at the between level, will keep the number of parameters to estimate between 7 and 14 (3 loading, 4 error variances, 4 thresholds, and 1 factor variance). Also, the scale of the observed categorical data will consist of four response categories, not exactly like YRBS, but I want the variables to be treated as categorical not continuous.

Figure 2.6

Study Two/Three Model



Along with deciding which SEM model type or complexity, parameters must be given population values for which the sample is generated, which means deciding on the population values for the factor loadings, correlations, and variances. Muthén (2002) generated data with factor loadings of .8 (which is high), a factor variance of 1, correlation of .25, and residual variances of .36 in his two factor CFA model. These numbers were chosen because the reliability for each factor indicator would be .64 or roughly sixty-four percent. Reliability describes consistency of the measurement. For a CFA, it describes the internal consistency of the indicators used to measure a factor. Bollen (1989) describes reliability as part of the measure that is free of random error. Muthén and Muthén (2002) define reliability as the ratio of variance of factor indicator to total variance. The percentage of variance explained by the factor for the particular indicator/ indicator reliability is given by

$$\text{Reliability=}\frac{\lambda^2\Psi}{\lambda^2\Psi+\theta} \tag{57}$$

where λ is the factor loading, Ψ is the factor variance, and θ is the residual variance (Muthén & Muthén, 2002). Fornell and Larcker (1981) states when reliability is less than .50, variance due to measurement error (i.e., the unreliable part) is larger than the variance captured by the construct. Muthén (2002) used a reliability measure of .64 when generating data for normal and non-normal SEM models for sample size recommendations. For the multilevel model, Muthén (1991) describes reliability as having a within and between measure, namely,

$$\text{Within Reliability } (y_g) = \frac{WF}{WF+WE} \tag{58}$$

$$\text{Between Reliability } (y_g)=\frac{BF}{BF+BE} \tag{59}$$

where BF and WF are within and between factor variances, and BE and WE are the measurement errors for the between factor and within factor, respectively. For a random effects CFA model, the correlation between two individuals within a group g for a variable y, is called the ICC (Muthén, 1991).

$$\text{ICC=}\ \frac{Cov(y_{gi},y_{gk})}{\sigma_{y_g}}=\frac{BF+BE}{(BF+BE)+(WF+WE)} \tag{60}$$

$$\text{True ICC or Latent variable ICC=}\frac{BF}{BF+WF} \tag{61}$$

The true ICC assumes error free variance. This essentially says the ICC can be approximated by expressing it as function of model parameters (Julian, 2001).

$$ICC = \frac{\lambda_b{}^2\psi_b + \theta_b}{(\lambda_b{}^2\psi_b + \theta_b) + \lambda_w{}^2\psi_w + \theta_w} \tag{62}$$

The symbol λ stands for the factor loading, Ψ for factor variance, and θ is the residual variance for the between and within factors. In order to calculate the latent ICC for the latent variable, we assume cross level invariance (i.e., the two constructs are being measured in the same way at the between or within level, and the factor loadings must be equated across levels; Selig et al., 2008). It should also be noted that the latent factor or one of the factor loadings has to be fixed since factors have no scale. Also, it should be noted that measurement error attenuates the factor ICC. Therefore, the approximated observed ICC will always be less than the factor ICC. The reliability used for this study will be similar to Muthén’s (2002) study, roughly .64. Using Muthén's (2002) formula for estimation of the observed ICC (which similar to Julian's (2001) study for the effect of ignoring clustering), I wanted to use three ICC conditions (low, medium, and high). However, once I generated the data, the original values changed due to the creation of data for two class populations and analyses of it for one class non-normal population. There were too many variables to control. Because this is the first study of its kind, a simpler route is preferred; as such the ICC will not vary. The estimated ICC for the moderate skew case was .20 and .09 for the severely skewed data.

$$ICC_M = \frac{\lambda_b{}^2\psi_b + \theta_b}{(\lambda_b{}^2\psi_b + \theta_b) + \lambda_w{}^2\psi_w + \theta_w} = \frac{(.80)^2(1) + (.36)}{((.80)^2(1) + (.36)) + ((1.58)^2 + 1.404} = .20 \tag{63}$$

Estimation method. Hox et al. (2010) compared five estimation methods along with several other dependent variables for determining sample size requirements, namely MUML, Muthén’s psuedobalanced method, which assumes all groups are balanced, ignores unbalance, Full Maximum Likelihood (ML), Robust Maximum likelihood (MLR), weighted least square mean adjusted (WLSM), and weighted least square mean and variance adjusted (WLSMV). He found that for model fit the MLR and WLSM Chi squared value was better than the MUML, but not as good as ML and WLSMV. He found that for standard error bias of the parameter, maximum likelihood (ML), WLSM, and WLSMV all had good coverage. However, MLR did not perform well for a small number of groups, and variance for all methods did not perform well with ML, giving acceptable CI for groups of sample size of 200. He noted the pseudo-balanced method had a negative effect (downward bias) on both standard errors and on the chi-squares, leading Hox not to recommend this method. MUML is a limited maximum likelihood method implemented in Mplus that ignores balanced groups and only works with random intercept models. ML and MLR are full maximum likelihood estimators that allow for random slopes and do not require groups to be balanced. MLR is the robust ML estimation that performs better than ML under conditions of moderate non-normality but requires a large sample size (Hox et al., 2010). WLSM and WLSMV are the limited information, diagonal weighted, least square estimated in Mplus. They are robust methods and do not perform as well as ML under multivariate normality (Hox et al., 2010). Mplus does not implement MUML for multilevel (type = twolevel) data that is has at least one observed data that is categorical. After excluding MUML, seven estimators are available in Mplus that could be used in my studies. Because of the limited scope of this study and known properties of these estimates, I am limiting the estimators in this study to four estimators: ML, MLR, WLSMV, and WLSM. The excluded estimators will not be included due to the research pointing to their shortcomings. For example, weighted least square estimator (WLS) is available for multilevel Mplus models. However, for small sample sizes, it is known that WLS is not as efficient as WLSM and WLSMV, which are robust WLS estimators that use the diagonal matrix instead of the full matrix. Therefore, four estimators will be used to estimate required sample size for non-normal continuous and non-normal categorical multilevel SEM data. Finnery & DiStefano (2006) recommended ML for moderate continuous non-normal data, and WLSMV for moderate or severe categorical non-normal data.

Software. What has been proposed to be varied so far has been the state of balance, sample size, group size, and estimation method. Since this study is also studying the impact of non-normality and impact of having empty cells, another variable emerges, degree of non-normality. In the previous proposal that started this investigation, the data from the national survey were ordinal (with ceiling/floor effect), and the data were highly skewed (with most students answering affirmative to one category). Nothing that educational data almost always have dependencies and survey data are rarely continuous, this study wants to investigate how those conditions impact sample size requirements for the various estimation methods in multilevel SEM.

For study two, continuous non-normal data has to be generated. There were two methods I found that could possibly assist in generating this type of data. I could create and combine the within and between matrix in Mplus, export the data into SAS, and then generate mixed normal continuous data in SAS using Fleishman's power transformation method. The Fleishman method is a method for generating sample data to the desired skewness and kurtosis from normal variables (Fan et al., 2001). The Fleishman (1978) polynomial transformation formula is given by:

$$Y = a + bZ + cZ^2 + dZ^3 \tag{64}$$

Where Y is the transformed non-normal variable, and a, b, c, d are coefficients needed to transform the normal variable to a non-normal variable. The transformation of the normal data using Fleishman (1978) formula will result in a positively skewed distribution. Manipulation of the formula can allow for negative skew (as cited in Fan et al., 2001). This is a great method, but it is a brute force method that created too many possibilities for human error and was time consuming. A rarely used method where two distributions were created and combined to create one non-normal distribution was developed by Muthén in Mplus. This method was found in Muthén's (2002) paper on sample size and power requirement. This method has not been used on multilevel data, and there was less guidance on how to utilize it to create the non-normal data as opposed to the SAS method. However, this method, if implemented correctly, is less messy. Therefore, I decided to use Mplus 6.1 to generate data for both of my studies. However, since I was doing so for many conditions and only had one computer, I had to use Mplus on Linux and write a script to use multiple nodes (multithreading techniques) to run multiple simulations at one time.

Non-Normality/Sparseness. When deciding the degree of non-normality, one has to start with the definition. There are two definition of normal using skew and kurtosis. Tukey (1960) defined normality based on using lambda (***λ***) as a distribution with skewness and kurtosis that were both zero (as cited in Fan et al., 2001). However, Ramberg et al. (1979) expounded on the work of Tukey, defining skewness as the third moment about the mean and kurtosis as the fourth moment (Fan, 2001, p. 64):

$$\alpha_3 = \frac{E(X-\mu)^3}{\sigma^3} \tag{65}$$

$$\alpha_4 = \frac{E(X-\mu)^4}{\sigma^4} \tag{66}$$

However, there is more than one definition of kurtosis. Excess kurtosis is defined as

$$\alpha_4 - 3 = \frac{E(X-\mu)^4}{\sigma^4} - 3 \tag{67}$$

Using this definition of kurtosis, the second definition of normality states that a distribution is normal if it has skewness of zero and kurtosis of three. Finnery and DiStefano (2006) defines moderate non-normality as having a skew less than 2, kurtosis less than 7, and severe non-normality as having a skew more than 2 and kurtosis greater than 7. My moderate non-normality values was estimated from the population had skew and kurtosis of 1.21 and 3.29(respectively). My severe non-normal skew and kurtosis was 2.29 and 7.465 (respectively). The data section has more detail in this section and has more detail of the data generation.

The third study of non-normal categorical data will also be generated and analyzed in Mplus 6.1. Mplus has several methods for generating non-normal categorical data. I used the threshold method that uses cut points to define the percentage of data that falls in each of the categories (four response categories). Thresholds depend on the estimator used. ML based estimators are based on the logit link function or logistic curve. In order to obtain the cut points, I have to obtain the logistic quartiles for the cumulative areas. WLSM and WLSMV estimators uses a probit as its' link function that is based on a cumulative Z distribution to define each category. Because my goal is to simulate the national data set that had floor and ceiling effects (very little variability), I will have some categories with very little data or sparse data. For four point scale (k), there is k-1 or three thresholds that have to be defined. Study 3 will have two conditions measuring sparseness. Sparseness 1: 90% of the data will fall into Category 1, 5% Category 2, 3% in Category 3, and 2% in the subsequent category. If we are using the cumulative Z distribution (standard normal distribution) when 90% of the data are in the first category, the threshold or z value is 1.28. If 5% of the data are in the next category (which means 95% cumulatively), the threshold or z value is 1.645. Sparseness condition two will mimic 50% of the data in the first category, 30% in the second, and the subsequent categories will have 10% of the data. The goal in Study 3 was to examine how sparseness affects the sample size requirement for the various estimators when there is 2%, 3%, and 5% degree of sparseness.

Data. The code to come up with study 2 was very difficult. As I stated before, I originally considered creating mixed non-normal continuous data in SAS using the Fan and Fan (2005) article as a framework to build a multilevel CFA. After finding it difficult to model the between structure and create dependencies in the SEM multilevel in SAS, I came across the Muthén & Muthén (2002) article that was the only article showing how to create a mixture distribution in Mplus using a mixture of two normal distributions. This method of making non-normal data was rarely used in Mplus according to my research. I combined the Muthén and Muthén (2002) code with the Hox and Maas (2010) code that showed how one could make a multilevel model in Mplus. The combined codes create a non-normal continuous distribution that is multilevel.

The mixture distribution is created by generating two classes and analyzing it as though it was one class. Muthén (2002) starts doing this by specifying the number of people in each class. Class one is referred to as the minority or outlier class in the paper, and [c#1@-2] tells how many people are associated in the class by translating the logit of -2 to a probability,

$$P = \frac{1}{1+e^{-L}} = \frac{1}{1+e^{-2}} = .12 \tag{68}$$

This probability indicates that 12% are in the minority class and 88% are in the majority class. The mixture of two normal distributions with different means and standard deviations is then used to create one mixed normal distribution. After defining the number of people in the class, the different factor means and standard deviations were defined for each class in the model generation portion of the Monte Carlo code. In Muthén's (2002) example the minority class, class one, had a factor mean of 15 and variance of 5 while the majority class was standard normal. The result was a population with skew of 1.2 and kurtosis from 1.5 to 1.6 for indicators y6 to y10 (or for the factor in which a mixture distribution was comprised). The choice of proportions, different means, and variances all affect the skewness and kurtosis. Once you decide on the proportions, one can simply manipulate the mean and variances to obtain the skew and kurtosis wanted.

To approximate the population values for my model, skew, and kurtosis (after deciding the means and standard deviations) step two is to run one replication with a large sample size (one million for mine) for the model you have chosen. You run this model to obtain the actual population values for the one class model. After running your proposed model on a large sample (by looking at the estimate average) you obtain the correct intercepts, and the within factor loading for the non-normal or mixed normal population. The third and last step, according to Muthén (2002), is to then use the new parameter estimates for the factor loadings to solve for the residual variances using the factor indicator reliabilities set by the user.

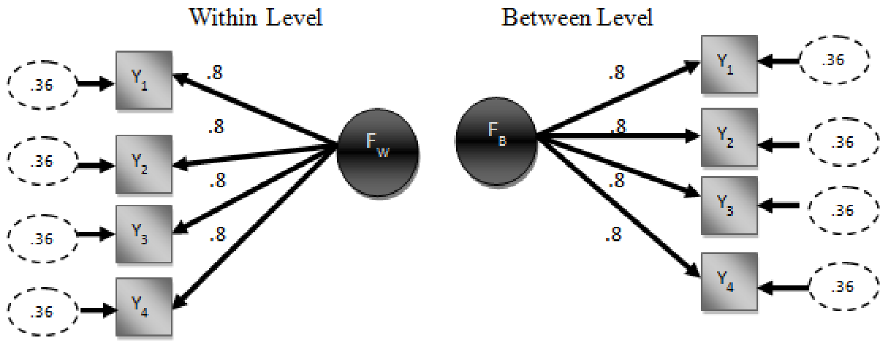
Using .8 as my loadings for the between and within levels, one as my factor variances and .36 as my error variances, I was able to obtain item reliabilities of .64 for both levels, see figure 23.

$$\mathrm{Re}\,liability = \frac{\lambda^2\psi}{\lambda^2\psi + \theta} = \frac{(.8)^2(1)}{(.8)^2(1) + .36} = .64 \tag{69}$$

I ran a large sample one replication model with varying different mean and standard deviation pairings until I got the skew and standard deviation desired. Using Muthén's (2002) code as a guideline, my portions were also 12% for the minority class and 88% for the majority. I also allowed one class to be standard normal and manipulated the mean and kurtosis for the minority class.

Figure 2.7

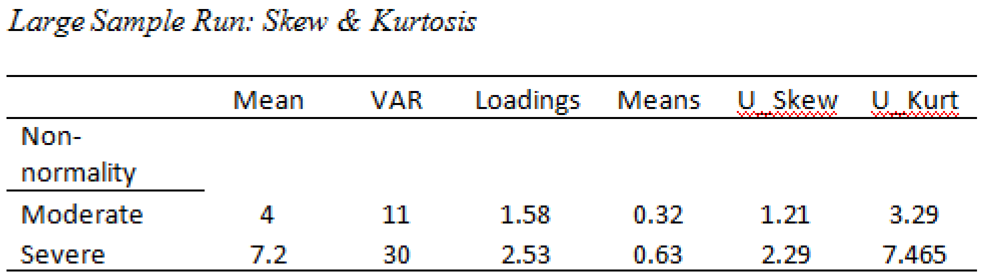
Initial model for large sample run



I used Lei and Lomax (2005) and the Finney and Distefano (2006) as guidelines on where to set my distributional values for moderate and severely non-normal conditions. Lei and Lomax (2005), using the definition of normality of zero skew and zero kurtosis, defined slight non-noramlity as skewness between .3 and .4, with kurtosis around 1.0. Severe non-normality was skewness above .7 and kurtosis above 3.5. Finney and Distefano (2006) defined moderate non-normality as having skew below 2 and kurtosis below 7, with sever non-normality being above those numbers respectively for both measures. Finney and Distefano (2006) used the second definition of normality based on skew of zero and kurtosis of three (i.e., excess kurtosis).

The population values for moderate non-normality were based on the minority class having mean 4 and variance 11, and on severe non-normality having mean 7.2 and variance 30. The majority class was held at standard normal. As table 2.1 shows, when the minority class was N(4,11), the loadings went from .8 to an average of 1.58 for the moderate non-normal CFA model. The data generated had a univariate skew of 1.21 and kurtosis 3.29. The skew and kurtosis values were found using excel, which define normality as N(0,3) and uses excess kurtosis for the definition of kurtosis. Severe non-normality was generated when the minority class was N(7.2,30). The within loading increased from .8 to 2.53, on average, and the intercept was .63. The severe non-normality had univariate skew of 2.29 and kurtosis of 7.465.

Table 2.1



Large Sample Run: Skew & Kurtosis

	Mean	VAR	Loadings	Means	U_Skew	U_Kurt
Non-normality						
Moderate	4	11	1.58	0.32	1.21	3.29
Severe	7.2	30	2.53	0.63	2.29	7.465

Muthén's (2002) mixed two normal classes N(0,1) and N(15,5) to obtain a mixture distribution. The original loading went from originally being .8 with .36 error variance and a reliability of .64, to having a loading of 4 and means of 1.42 for the one class non-normal distribution. He then used this information about the loading, and for a set reliability of .64, he solved for the error variance of 9. Similarly, after obtaining a loading of 1.58, I decided to use 1.404 as my error variance to keep the reliability to .64 for the moderate non-normal case. For the severe non-normal case, I decided to use 3.6005 as my error variance which gave me an item reliability of .64.

Assessing Sample Size. Muthén and Muthén (2002), which outlines how to determine sample size and power, is one of a number of methods to determine sample size. Hoogland and Boosma (1998) outlined three criteria for what they assess to be sufficient sample size: relative parameter bias of no more than 5%, relative bias of the standard errors of no more than 10%, and the chi-square statistic rejection rate at 5%. Starting with version 3.0, Mplus began summarizing fit statistics that allowed researchers to compare at the critical chi-square for the given degrees of freedom at various alpha levels (via the expected percentiles or critical value and expected proportion column or alpha) to the observed chi-square test statistic and observed alpha from the monte carlo replication (via observed percentile and observed proportion, respectively). If the observed percentile or proportion is bigger than the expected or theoretical chi-square values, then the chi-square distribution is said to not to be well approximated. Note, assessing chi-square statistics solely at the .05 rate is unreliable (Hoogland and Boosma, 1998). The chi-square relative bias (Bandalos, 2006; Brown, 2006), is another method of determining if the chi-square distribution and is well approximated:

$$Bias(\overline{\chi^2}) = \frac{\overline{\chi^2} - df}{df} \tag{70}$$

where $\overline{\chi^2}$ is the average chi-square statistic across all replications, and df is the degrees of freedom. Instead of looking at the percentage of sample chi-square values that fall in the rejection region as being equal to the nominal alpha rate, .05, Bradley (1978) suggests comparing the criterion for the type one area rate lying in the range $\alpha \pm .1\alpha$ or $\alpha \pm .5\alpha$ which are between (.045,.055) or (.025, .075) to an alpha of .05 (as cited in Bandalos, 2006). The mean of the chi-square distribution is its degrees of freedom. The percentage of chi-square that fall in the rejection region is the percentage of time the null was rejected (where the null is $H_o : \mu = df$ and the alternative is $H_a : \mu \neq df$). The percentage of time the null is rejected when it is true is known as the proportion of time one makes a type one error. For some estimators (e.g. WLS based estimators and MLR) the chi-square df is not the true chi square value. Because of this fact, the proportion of rejections for which the critical value is exceeded will be used to judge if the chi-square distribution was well approximated. The first Bradley (1978) confidence interval is very conservative, whereas, the second one is liberal. In this study the liberal confidence interval will be utilized.

Summary. Study 2 will have 2 balanced/unbalanced dichotomous conditions, 3 cluster sizes, 4 estimation methods, 3 sample sizes, 2 non-normality conditions with 1000 reps per condition. Therefore 2x 3 x 4 x 3 x 2 yields 144 total conditions. Study 3 will have 4 estimation methods, 3 sample sizes, 3 balanced and 3 unbalanced group sizes, and 2 non-normality conditions, with 1000 reps per condition and a total of 144 conditions.

Limitations. As with most simulations, you can't include every possibility. The scope of the study is limited in that no missing data are assumed, there are not structural relationships taken into account, and the ICC is not varied. My model is a very simple model. Future studies need to explore missing data or if the pattern of missing data has an effect on the sample size recommendation or even if model misspecification has an effect on sample the size recommendation for between level MSEM. The most important test analyze if varying the ICC has an impact on the sample size recommendation.

CHAPTER FOUR

STUDIES TWO AND THREE RESULTS

This section presents the results of the study of sample the size requirement for a multilevel structural equation model (SEM) under non-normality (namely, studies 2 and 3). Study two studies the performance of estimators for non-normal continuous data, while study three analyzes the performance of estimators for categorical data to determine sample the size requirement for a multilevel SEM model. The subsequent chapter, chapter five, provides a deeper discussion of the results presented here and the formal conclusion drawn. The results are presented in linear numerical order (i.e. study two's results will be followed by study three's results).

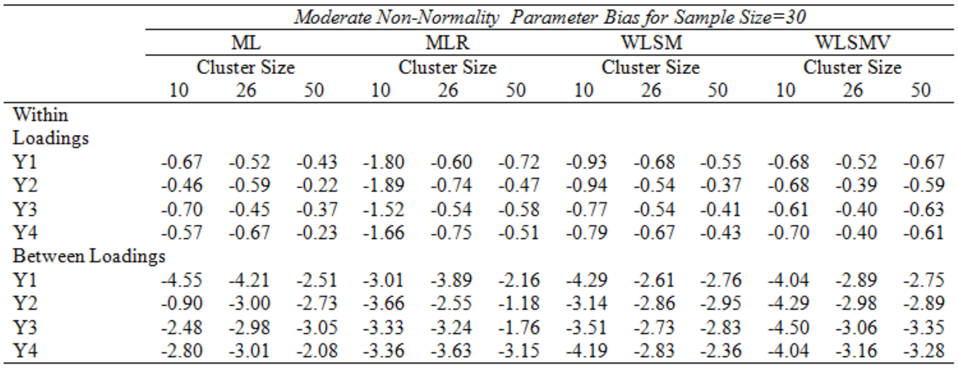
There are a number of methods to assess the appropriate sample size requirement for a given estimator. Muthén and Muthén (2002) gave three requirements for determining the best samples size: parameter bias does not exceed 10%, standard error bias does not exceed 10%, and coverage range between .91 and .98 (Brown, 2006). Hoogland and Boosma (1998) outline the criteria for sufficient sample size when assessing the performance of estimators. Their criteria were a relative parameter bias of no more than 5%, a relative bias of the standard errors of no more than 10%, and a rejection rate for the chi-square statistic at the 5% nominal alpha level. Meuleman and Billet (2009) augmented the Hoogland and Boosma (1998) criteria by also requiring coverage of .95 or better, and expecting researchers to look at the number of inadmissible solutions. Muthén and Muthén’s (2002) recommendations are the primary guideline for assessing sample size. The secondary guideline for assessing sample size will be the rejection rate for the chi-square statistic and the number of inadmissible solutions.

Results of Study Two

Parameter Bias. The average parameter bias should not exceed 10% for any of our four estimators, namely, maximum likelihood (ML), robust ML (MLR), weighted least square mean adjusted (WLSM), and weighted least square mean and variance adjusted (WLSMV). Recall, two conditions were moderately non-normal and severely non-normal for balanced and non-balanced data.

Table 2.2

Moderate Non-Normality Parameter Bias for Sample Size=30



	Moderate Non-Normality Parameter Bias for Sample Size=30											
	ML			MLR			WLSM			WLSMV		
	Cluster Size			Cluster Size			Cluster Size			Cluster Size		
	10	26	50	10	26	50	10	26	50	10	26	50
Within Loadings												
Y1	-0.67	-0.52	-0.43	-1.80	-0.60	-0.72	-0.93	-0.68	-0.55	-0.68	-0.52	-0.67
Y2	-0.46	-0.59	-0.22	-1.89	-0.74	-0.47	-0.94	-0.54	-0.37	-0.68	-0.39	-0.59
Y3	-0.70	-0.45	-0.37	-1.52	-0.54	-0.58	-0.77	-0.54	-0.41	-0.61	-0.40	-0.63
Y4	-0.57	-0.67	-0.23	-1.66	-0.75	-0.51	-0.79	-0.67	-0.43	-0.70	-0.40	-0.61
Between Loadings												
Y1	-4.55	-4.21	-2.51	-3.01	-3.89	-2.16	-4.29	-2.61	-2.76	-4.04	-2.89	-2.75
Y2	-0.90	-3.00	-2.73	-3.66	-2.55	-1.18	-3.14	-2.86	-2.95	-4.29	-2.98	-2.89
Y3	-2.48	-2.98	-3.05	-3.33	-3.24	-1.76	-3.51	-2.73	-2.83	-4.50	-3.06	-3.35
Y4	-2.80	-3.01	-2.08	-3.36	-3.63	-3.15	-4.19	-2.83	-2.36	-4.04	-3.16	-3.28

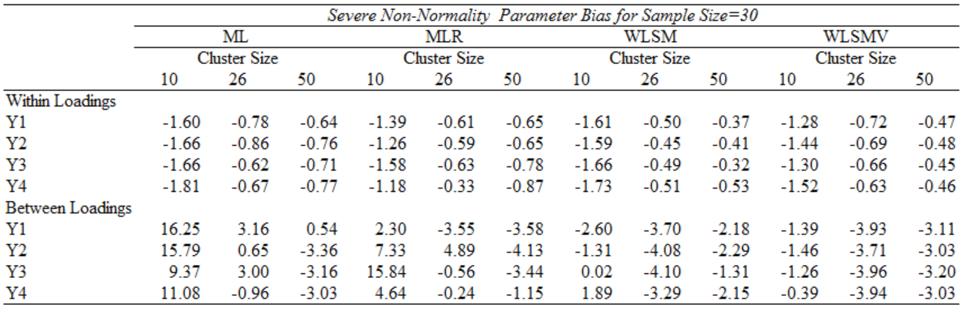
For a sample size or the number of clusters equaling 30, there was no large parameter bias. On average, the biases were all less than 5% for each of the estimators used in the study under the condition of moderate non-normality and balanced sample size.

When the sample size is 30 and we look at parameter bias under the condition of severe non-normality with balanced sample size there is large parameter bias at the between level for ML based estimators when the number of subjects within the clusters or cluster size is ten (see

Table 2.3). In order to better understand the pattern of behavior across conditions, the average of the parameter biases for all the indicators, y1-y4, was also tabulated and compared.

Table 2.3

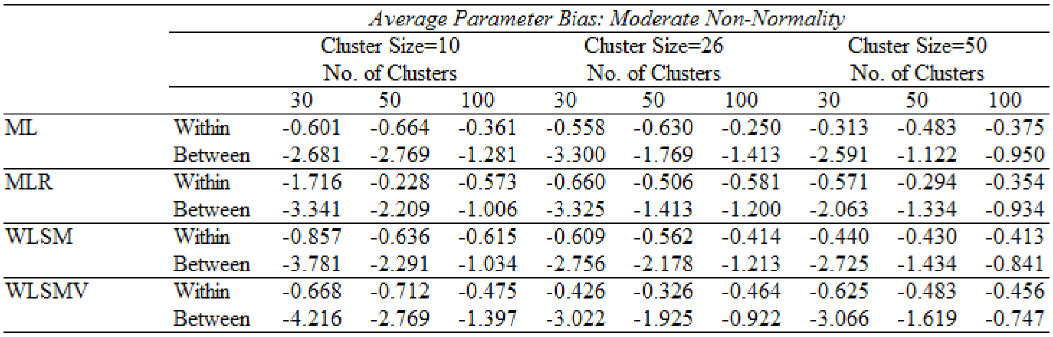
Severe Non-Normality Parameter Bias for Sample Size=30



	Severe Non-Normality Parameter Bias for Sample Size=30											
	ML			MLR			WLSM			WLSMV		
	Cluster Size			Cluster Size			Cluster Size			Cluster Size		
	10	26	50	10	26	50	10	26	50	10	26	50
Within Loadings												
Y1	-1.60	-0.78	-0.64	-1.39	-0.61	-0.65	-1.61	-0.50	-0.37	-1.28	-0.72	-0.47
Y2	-1.66	-0.86	-0.76	-1.26	-0.59	-0.65	-1.59	-0.45	-0.41	-1.44	-0.69	-0.48
Y3	-1.66	-0.62	-0.71	-1.58	-0.63	-0.78	-1.66	-0.49	-0.32	-1.30	-0.66	-0.45
Y4	-1.81	-0.67	-0.77	-1.18	-0.33	-0.87	-1.73	-0.51	-0.53	-1.52	-0.63	-0.46
Between Loadings												
Y1	16.25	3.16	0.54	2.30	-3.55	-3.58	-2.60	-3.70	-2.18	-1.39	-3.93	-3.11
Y2	15.79	0.65	-3.36	7.33	4.89	-4.13	-1.31	-4.08	-2.29	-1.46	-3.71	-3.03
Y3	9.37	3.00	-3.16	15.84	-0.56	-3.44	0.02	-4.10	-1.31	-1.26	-3.96	-3.20
Y4	11.08	-0.96	-3.03	4.64	-0.24	-1.15	1.89	-3.29	-2.15	-0.39	-3.94	-3.03

Table 2.4

Average Parameter Bias: Moderate Non-Normality

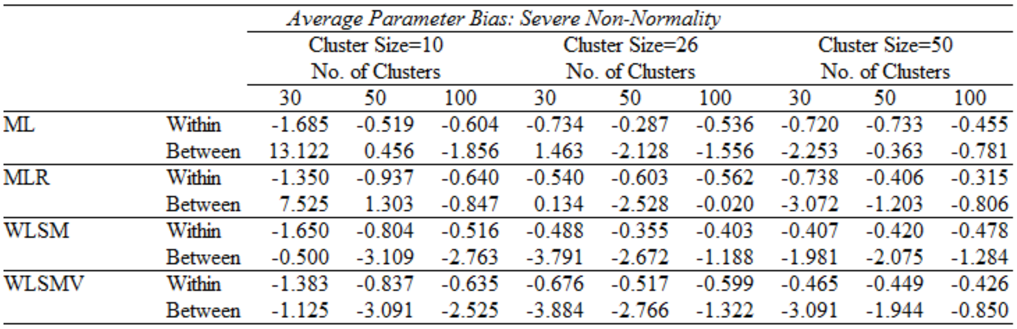


		Average Parameter Bias: Moderate Non-Normality								
		Cluster Size=10			Cluster Size=26			Cluster Size=50		
		No. of Clusters			No. of Clusters			No. of Clusters		
		30	50	100	30	50	100	30	50	100
ML	Within	-0.601	-0.664	-0.361	-0.558	-0.630	-0.250	-0.313	-0.483	-0.375
	Between	-2.681	-2.769	-1.281	-3.300	-1.769	-1.413	-2.591	-1.122	-0.950
MLR	Within	-1.716	-0.228	-0.573	-0.660	-0.506	-0.581	-0.571	-0.294	-0.354
	Between	-3.341	-2.209	-1.006	-3.325	-1.413	-1.200	-2.063	-1.334	-0.934
WLSM	Within	-0.857	-0.636	-0.615	-0.609	-0.562	-0.414	-0.440	-0.430	-0.413
	Between	-3.781	-2.291	-1.034	-2.756	-2.178	-1.213	-2.725	-1.434	-0.841
WLSMV	Within	-0.668	-0.712	-0.475	-0.426	-0.326	-0.464	-0.625	-0.483	-0.456
	Between	-4.216	-2.769	-1.397	-3.022	-1.925	-0.922	-3.066	-1.619	-0.747

The average parameter bias across the parameters, y1-y4, shows that the bias tended to decrease as sample size (number of clusters at the between level) increased. This pattern was observed only at the between level (for example, table IX, WLSMV -4.216, -2.769, -1.397 for cluster size 10, and sample size 30, 50 and ,100, respectively). Sometimes this trend was observed at the within level but due to sampling variation it is difficult to judge.

Table 2.5

Average Parameter Bias: Severe Non-Normality

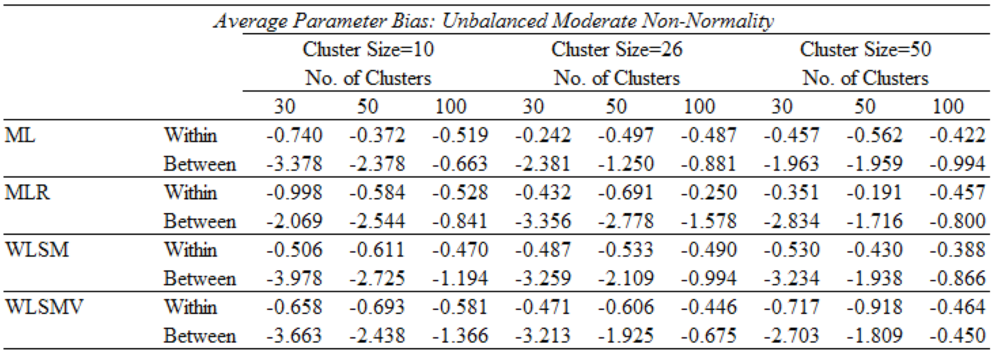


		Average Parameter Bias: Severe Non-Normality								
		Cluster Size=10			Cluster Size=26			Cluster Size=50		
		No. of Clusters			No. of Clusters			No. of Clusters		
		30	50	100	30	50	100	30	50	100
ML	Within	-1.685	-0.519	-0.604	-0.734	-0.287	-0.536	-0.720	-0.733	-0.455
	Between	13.122	0.456	-1.856	1.463	-2.128	-1.556	-2.253	-0.363	-0.781
MLR	Within	-1.350	-0.937	-0.640	-0.540	-0.603	-0.562	-0.738	-0.406	-0.315
	Between	7.525	1.303	-0.847	0.134	-2.528	-0.020	-3.072	-1.203	-0.806
WLSM	Within	-1.650	-0.804	-0.516	-0.488	-0.355	-0.403	-0.407	-0.420	-0.478
	Between	-0.500	-3.109	-2.763	-3.791	-2.672	-1.188	-1.981	-2.075	-1.284
WLSMV	Within	-1.383	-0.837	-0.635	-0.676	-0.517	-0.599	-0.465	-0.449	-0.426
	Between	-1.125	-3.091	-2.525	-3.884	-2.766	-1.322	-3.091	-1.944	-0.850

As seen in Table 2.5 for balanced severe non-normality, when the cluster size and sample size is low for the ML based estimators, the average between level parameter bias is very high with the ML condition surpassing the 10% bias mark (see Table 2.5, between level ML number of clusters 30 and cluster size is 10, 13.122% bias).

Table 2.6

Average Parameter Bias: Unbalanced Moderate Non-normality



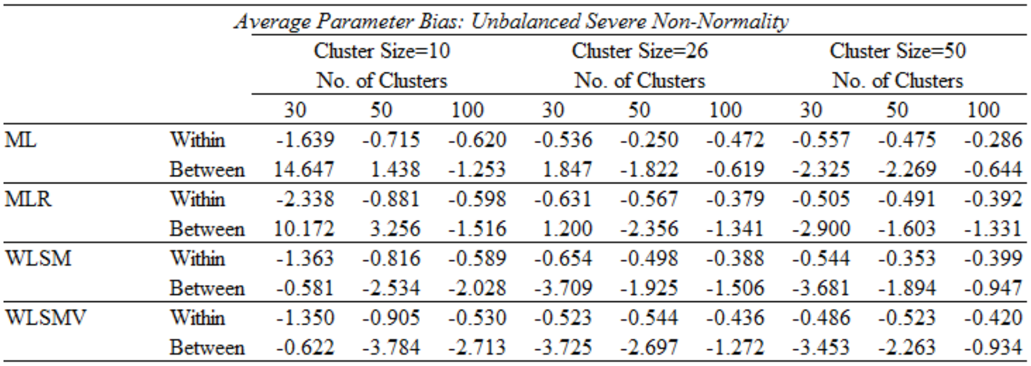
		Average Parameter Bias: Unbalanced Moderate Non-Normality								
		Cluster Size=10			Cluster Size=26			Cluster Size=50		
		No. of Clusters			No. of Clusters			No. of Clusters		
		30	50	100	30	50	100	30	50	100
ML	Within	-0.740	-0.372	-0.519	-0.242	-0.497	-0.487	-0.457	-0.562	-0.422
	Between	-3.378	-2.378	-0.663	-2.381	-1.250	-0.881	-1.963	-1.959	-0.994
MLR	Within	-0.998	-0.584	-0.528	-0.432	-0.691	-0.250	-0.351	-0.191	-0.457
	Between	-2.069	-2.544	-0.841	-3.356	-2.778	-1.578	-2.834	-1.716	-0.800
WLSM	Within	-0.506	-0.611	-0.470	-0.487	-0.533	-0.490	-0.530	-0.430	-0.388
	Between	-3.978	-2.725	-1.194	-3.259	-2.109	-0.994	-3.234	-1.938	-0.866
WLSMV	Within	-0.658	-0.693	-0.581	-0.471	-0.606	-0.446	-0.717	-0.918	-0.464
	Between	-3.663	-2.438	-1.366	-3.213	-1.925	-0.675	-2.703	-1.809	-0.450

Table 2.6 displays parameter bias for moderate normality when the clusters are unbalanced. Looking at the averages, there seems to be no difference between unbalanced and balanced conditions for parameter bias in moderate non-normal data. Again, as sample size

increases (number of clusters increased at the between level) the bias decrease (Table 2.6, WLSMV between -3.663 to -2.438 to -1.366 decreased as sample size increased from 30 to 50 to 100). The parameter bias at the within level seems to not really be affected by sample size (see Table 2.6, WLSMV). The within level values of -.658 to -.698 to -.681 decreased as sample size increased from 30 to 50 to 100. This could be due to sampling variance so statistical tests have to verify what we are seeing.

Table 2.7

Average Parameter Bias: Unbalanced Severe Non-Normality



Average Parameter Bias: Unbalanced Severe Non-Normality

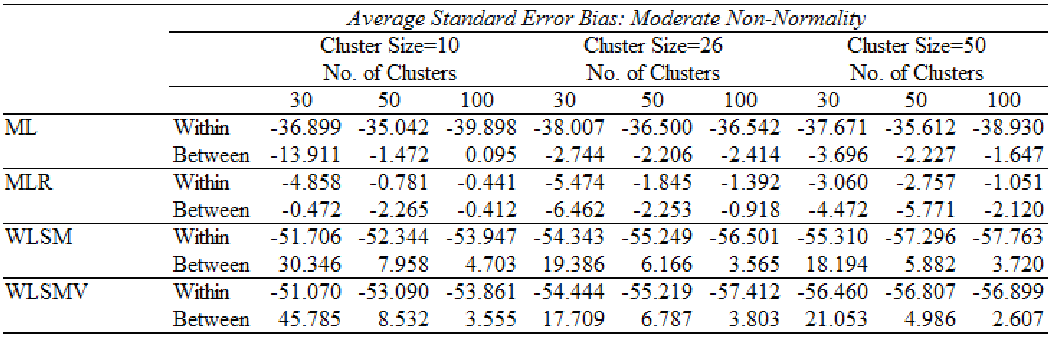
		Cluster Size=10			Cluster Size=26			Cluster Size=50		
		No. of Clusters			No. of Clusters			No. of Clusters		
		30	50	100	30	50	100	30	50	100
ML	Within	-1.639	-0.715	-0.620	-0.536	-0.250	-0.472	-0.557	-0.475	-0.286
	Between	14.647	1.438	-1.253	1.847	-1.822	-0.619	-2.325	-2.269	-0.644
MLR	Within	-2.338	-0.881	-0.598	-0.631	-0.567	-0.379	-0.505	-0.491	-0.392
	Between	10.172	3.256	-1.516	1.200	-2.356	-1.341	-2.900	-1.603	-1.331
WLSM	Within	-1.363	-0.816	-0.589	-0.654	-0.498	-0.388	-0.544	-0.353	-0.399
	Between	-0.581	-2.534	-2.028	-3.709	-1.925	-1.506	-3.681	-1.894	-0.947
WLSMV	Within	-1.350	-0.905	-0.530	-0.523	-0.544	-0.436	-0.486	-0.523	-0.420
	Between	-0.622	-3.784	-2.713	-3.725	-2.697	-1.272	-3.453	-2.263	-0.934

Similarly, to balance severe non-normality the ML based estimators had problems with small cluster size. Here, Table 2.7, both ML and MLR has a large positive between level parameter bias. Their bias is larger than what was seen for the balanced severe non-normality case (Table 2.5), 13.122 vs. 14.647 ML and 7.525 vs. 10.172 MLR (balanced vs. non-balanced, respectively). Again, other than the case where both the cluster size and sample size is small, there does not appear to be a difference between the balanced and the unbalanced case for parameter bias.

Standard Error Bias. According to Muthén and Muthén (2002) the standard error bias should not exceed 10%. Standard error bias is the difference between the average standard error and standard deviation divided by the standard deviation—multiplied by 100. Again, our four estimators are maximum likelihood (ML), robust ML (MLR), weighted least square mean adjusted (WLSM), and weighted least square mean and variance adjusted (WLSMV). Each of these are studied under the condition of moderate non-normality, and severe non-normality with balanced and non-balanced data.

Table 2.8

Average Standard Error Bias: Moderate Non-Normality

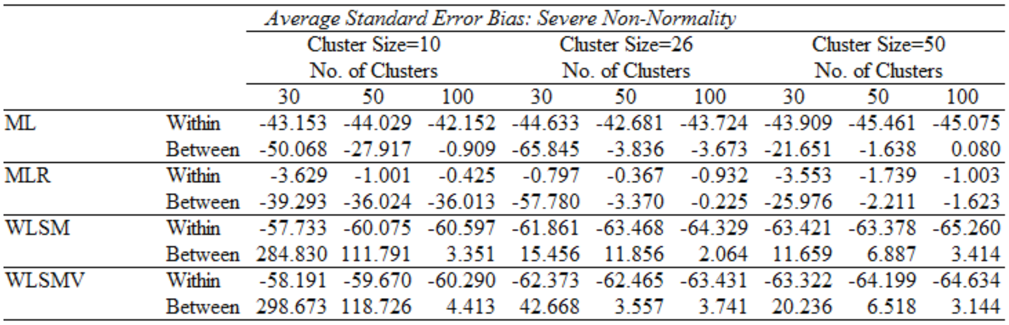


		Average Standard Error Bias: Moderate Non-Normality								
		Cluster Size=10			Cluster Size=26			Cluster Size=50		
		No. of Clusters			No. of Clusters			No. of Clusters		
		30	50	100	30	50	100	30	50	100
ML	Within	-36.899	-35.042	-39.898	-38.007	-36.500	-36.542	-37.671	-35.612	-38.930
	Between	-13.911	-1.472	0.095	-2.744	-2.206	-2.414	-3.696	-2.227	-1.647
MLR	Within	-4.858	-0.781	-0.441	-5.474	-1.845	-1.392	-3.060	-2.757	-1.051
	Between	-0.472	-2.265	-0.412	-6.462	-2.253	-0.918	-4.472	-5.771	-2.120
WLSM	Within	-51.706	-52.344	-53.947	-54.343	-55.249	-56.501	-55.310	-57.296	-57.763
	Between	30.346	7.958	4.703	19.386	6.166	3.565	18.194	5.882	3.720
WLSMV	Within	-51.070	-53.090	-53.861	-54.444	-55.219	-57.412	-56.460	-56.807	-56.899
	Between	45.785	8.532	3.555	17.709	6.787	3.803	21.053	4.986	2.607

For balanced, moderate, non-normal multilevel data, all estimators (except MLR) had large within level bias did not change or show a trend as cluster size or sample size (i.e. number of clusters) increased. However, the weighted least square (WLS) based estimators tended to have larger within level standard error bias (Table XIII). The standard error bias for MLR on moderately non-normal, continuous data seems to work well at the between and within levels, even for small sample size and cluster size (Table XIII). For ML, WLSM, and WLSMV, the within level standard error was larger than the 10% criterion, and there is a high standard error bias at the between level that coincides with a low sample size of thirty (WLSMV, 30/10, 45.785; 30/26, 17.709; and 30/50, 21.053).

Table 2.9

Average Standard Error Bias: Severe Non-normality



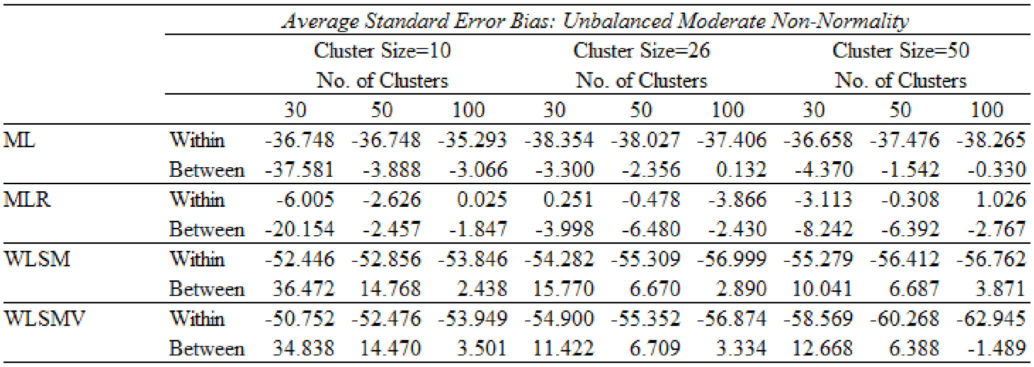
		Average Standard Error Bias: Severe Non-Normality								
		Cluster Size=10 No. of Clusters			Cluster Size=26 No. of Clusters			Cluster Size=50 No. of Clusters		
		30	50	100	30	50	100	30	50	100
ML	Within	-43.153	-44.029	-42.152	-44.633	-42.681	-43.724	-43.909	-45.461	-45.075
	Between	-50.068	-27.917	-0.909	-65.845	-3.836	-3.673	-21.651	-1.638	0.080
MLR	Within	-3.629	-1.001	-0.425	-0.797	-0.367	-0.932	-3.553	-1.739	-1.003
	Between	-39.293	-36.024	-36.013	-57.780	-3.370	-0.225	-25.976	-2.211	-1.623
WLSM	Within	-57.733	-60.075	-60.597	-61.861	-63.468	-64.329	-63.421	-63.378	-65.260
	Between	284.830	111.791	3.351	15.456	11.856	2.064	11.659	6.887	3.414
WLSMV	Within	-58.191	-59.670	-60.290	-62.373	-62.465	-63.431	-63.322	-64.199	-64.634
	Between	298.673	118.726	4.413	42.668	3.557	3.741	20.236	6.518	3.144

There was little standard error bias at the between or within levels, for MLR under moderate non-normality. However, under severe non-normality, the between level standard bias is very negative when the cluster size is low or when the sample size is low (Table 2.9, MLR 30/26,   -57.780; 30/50, -25.976). When severe non-normality is present, for WLS based estimators, the within level standard error bias is large and similar to the results under moderate non-normality. However, at the between level, there are huge positive between level biases that decrease as the sample size increases and as the cluster size increases (Table 2.9, WLSM 30/10, 284.830; 50/10, 111.791; and 30/26, 15.456). As sample size increased, for the ML estimator, the standard error bias at the between level decreases (30/10 -50.068 to 50/10 -27.917). That pattern did not hold as cluster size increased for a couple of these conditions for ML, but this could be due to sampling variance (ML 30/10, -50.068 to 30/26, -65.845). It is apparent that severe non-normality has an effect on standard error bias for all estimators, especially at the between level.

In Tables 2.10 and 2.11, below, the averaged standard error biases for the moderate and severe unbalanced non-normal data are shown. The data follows similar pattern as the balanced data; for low sample size there is generally high between level standard error bias. There does not seem to be a difference between standard error biases at the within level, but there is a slight difference at the between level (e.g. MLR balanced = 30/10 is -.472 & unbalanced = 30/10 is -20.154). This supports evidence that that the standard errors might be affected by unbalanced data.

Table 2.10

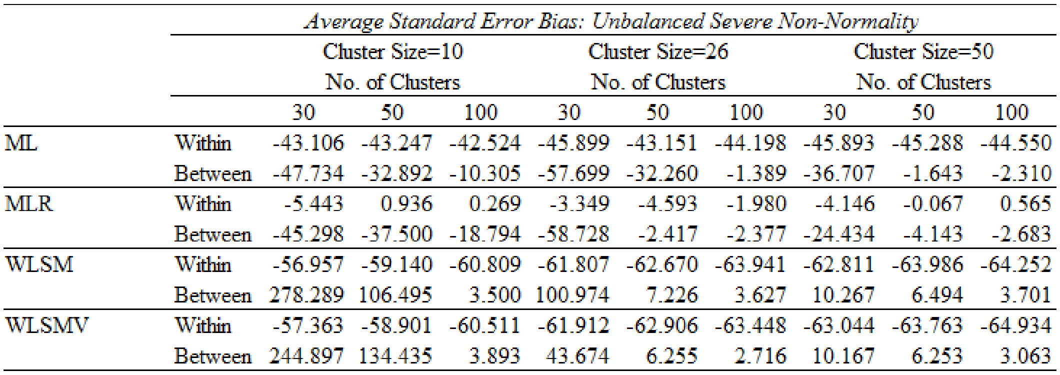
Average Standard Error Bias: Unbalanced Moderate Non-Normality



		Average Standard Error Bias: Unbalanced Moderate Non-Normality								
		Cluster Size=10			Cluster Size=26			Cluster Size=50		
		No. of Clusters			No. of Clusters			No. of Clusters		
		30	50	100	30	50	100	30	50	100
ML	Within	-36.748	-36.748	-35.293	-38.354	-38.027	-37.406	-36.658	-37.476	-38.265
	Between	-37.581	-3.888	-3.066	-3.300	-2.356	0.132	-4.370	-1.542	-0.330
MLR	Within	-6.005	-2.626	0.025	0.251	-0.478	-3.866	-3.113	-0.308	1.026
	Between	-20.154	-2.457	-1.847	-3.998	-6.480	-2.430	-8.242	-6.392	-2.767
WLSM	Within	-52.446	-52.856	-53.846	-54.282	-55.309	-56.999	-55.279	-56.412	-56.762
	Between	36.472	14.768	2.438	15.770	6.670	2.890	10.041	6.687	3.871
WLSMV	Within	-50.752	-52.476	-53.949	-54.900	-55.352	-56.874	-58.569	-60.268	-62.945
	Between	34.838	14.470	3.501	11.422	6.709	3.334	12.668	6.388	-1.489

Table 2.11

Average Standard Error Bias: Unbalanced Severe Non-Normality

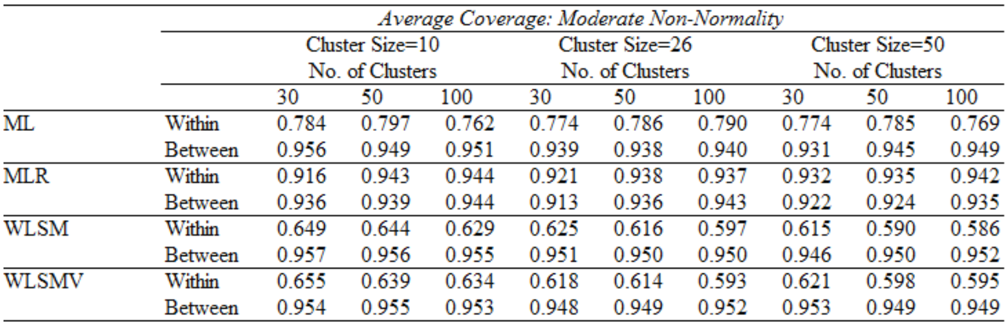


		Average Standard Error Bias: Unbalanced Severe Non-Normality								
		Cluster Size=10			Cluster Size=26			Cluster Size=50		
		No. of Clusters			No. of Clusters			No. of Clusters		
		30	50	100	30	50	100	30	50	100
ML	Within	-43.106	-43.247	-42.524	-45.899	-43.151	-44.198	-45.893	-45.288	-44.550
	Between	-47.734	-32.892	-10.305	-57.699	-32.260	-1.389	-36.707	-1.643	-2.310
MLR	Within	-5.443	0.936	0.269	-3.349	-4.593	-1.980	-4.146	-0.067	0.565
	Between	-45.298	-37.500	-18.794	-58.728	-2.417	-2.377	-24.434	-4.143	-2.683
WLSM	Within	-56.957	-59.140	-60.809	-61.807	-62.670	-63.941	-62.811	-63.986	-64.252
	Between	278.289	106.495	3.500	100.974	7.226	3.627	10.267	6.494	3.701
WLSMV	Within	-57.363	-58.901	-60.511	-61.912	-62.906	-63.448	-63.044	-63.763	-64.934
	Between	244.897	134.435	3.893	43.674	6.255	2.716	10.167	6.253	3.063

Coverage. Muthén and Muthén's (2002) last requirement specified that coverage be between .91 and .98 for assessing sample size requirements in structural equation modeling. Using this guide, I averaged the coverage across all indicators and obtained the average coverage.

Table 2.12

Average Coverage: Moderate Non-Normality



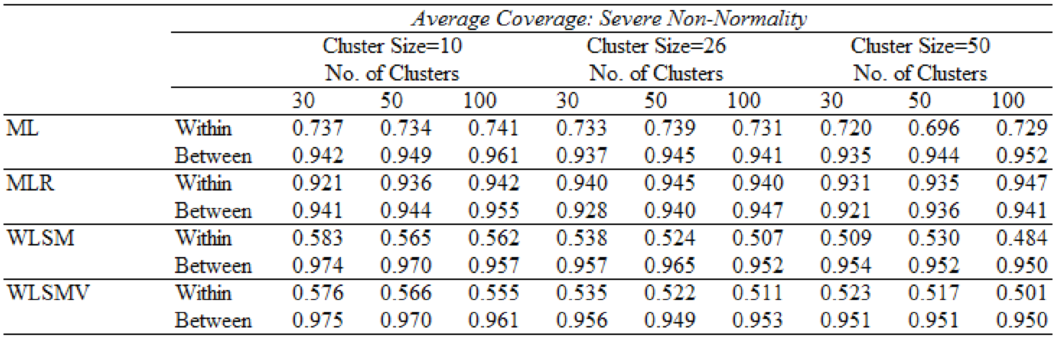
Average Coverage: Moderate Non-Normality

		Cluster Size=10 No. of Clusters			Cluster Size=26 No. of Clusters			Cluster Size=50 No. of Clusters		
		30	50	100	30	50	100	30	50	100
ML	Within	0.784	0.797	0.762	0.774	0.786	0.790	0.774	0.785	0.769
	Between	0.956	0.949	0.951	0.939	0.938	0.940	0.931	0.945	0.949
MLR	Within	0.916	0.943	0.944	0.921	0.938	0.937	0.932	0.935	0.942
	Between	0.936	0.939	0.944	0.913	0.936	0.943	0.922	0.924	0.935
WLSM	Within	0.649	0.644	0.629	0.625	0.616	0.597	0.615	0.590	0.586
	Between	0.957	0.956	0.955	0.951	0.950	0.950	0.946	0.950	0.952
WLSMV	Within	0.655	0.639	0.634	0.618	0.614	0.593	0.621	0.598	0.595
	Between	0.954	0.955	0.953	0.948	0.949	0.952	0.953	0.949	0.949

From Table 2.12, the within level coverage was poor for all estimators except MLR, and the between level was fine for all estimators in the moderate and severe non-normality case.

Table 2.13

Average Coverage: Severe Non-Normality



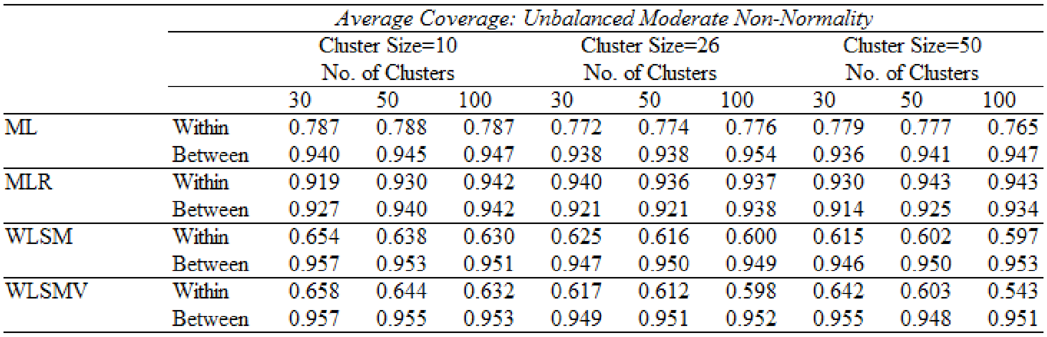
Average Coverage: Severe Non-Normality

		Cluster Size=10 No. of Clusters			Cluster Size=26 No. of Clusters			Cluster Size=50 No. of Clusters		
		30	50	100	30	50	100	30	50	100
ML	Within	0.737	0.734	0.741	0.733	0.739	0.731	0.720	0.696	0.729
	Between	0.942	0.949	0.961	0.937	0.945	0.941	0.935	0.944	0.952
MLR	Within	0.921	0.936	0.942	0.940	0.945	0.940	0.931	0.935	0.947
	Between	0.941	0.944	0.955	0.928	0.940	0.947	0.921	0.936	0.941
WLSM	Within	0.583	0.565	0.562	0.538	0.524	0.507	0.509	0.530	0.484
	Between	0.974	0.970	0.957	0.957	0.965	0.952	0.954	0.952	0.950
WLSMV	Within	0.576	0.566	0.555	0.535	0.522	0.511	0.523	0.517	0.501
	Between	0.975	0.970	0.961	0.956	0.949	0.953	0.951	0.951	0.950

The unbalanced average coverage follows the same pattern as the balanced. MLR has superior coverage both at the between and within levels under both conditions of normality. The other estimators between levels are well below the recommended range and the between levels are sufficient. There doesn't seem to be a real difference in coverage between the balanced and unbalanced conditions.

Table 2.14

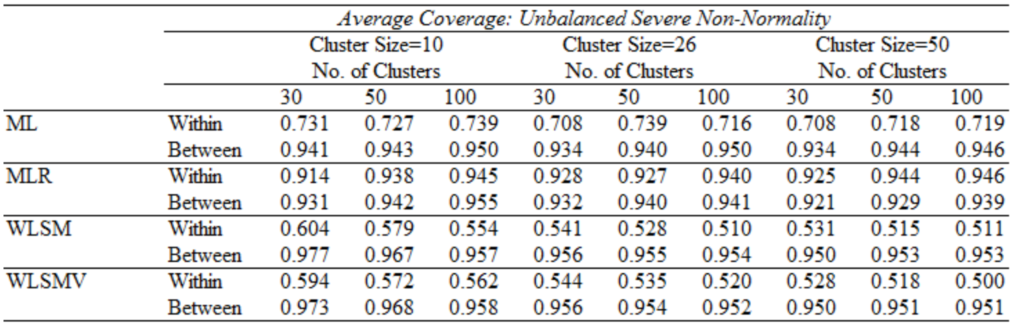
Average Coverage: Unbalanced Moderate Non-Normality



		Average Coverage: Unbalanced Moderate Non-Normality								
		Cluster Size=10			Cluster Size=26			Cluster Size=50		
		No. of Clusters			No. of Clusters			No. of Clusters		
		30	50	100	30	50	100	30	50	100
ML	Within	0.787	0.788	0.787	0.772	0.774	0.776	0.779	0.777	0.765
	Between	0.940	0.945	0.947	0.938	0.938	0.954	0.936	0.941	0.947
MLR	Within	0.919	0.930	0.942	0.940	0.936	0.937	0.930	0.943	0.943
	Between	0.927	0.940	0.942	0.921	0.921	0.938	0.914	0.925	0.934
WLSM	Within	0.654	0.638	0.630	0.625	0.616	0.600	0.615	0.602	0.597
	Between	0.957	0.953	0.951	0.947	0.950	0.949	0.946	0.950	0.953
WLSMV	Within	0.658	0.644	0.632	0.617	0.612	0.598	0.642	0.603	0.543
	Between	0.957	0.955	0.953	0.949	0.951	0.952	0.955	0.948	0.951

Table 2.15

Average Coverage: Unbalanced Severe Non-Normality

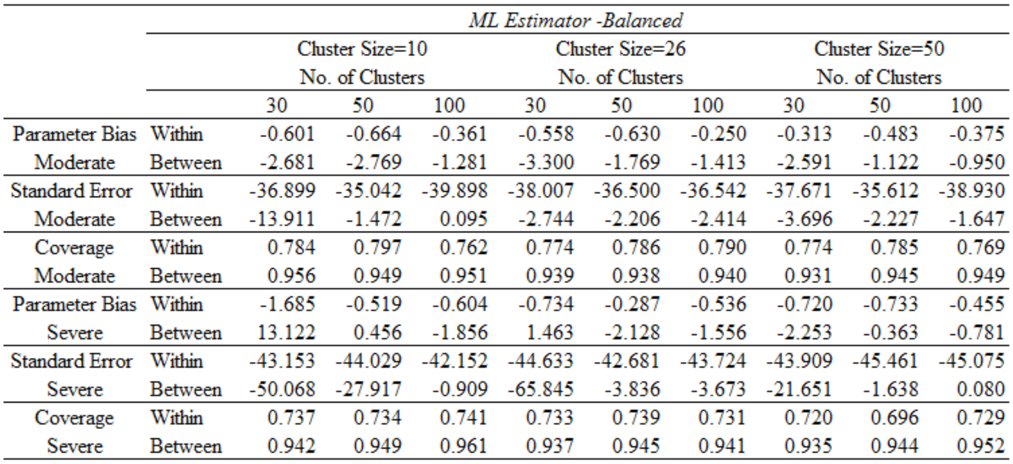


		Average Coverage: Unbalanced Severe Non-Normality								
		Cluster Size=10			Cluster Size=26			Cluster Size=50		
		No. of Clusters			No. of Clusters			No. of Clusters		
		30	50	100	30	50	100	30	50	100
ML	Within	0.731	0.727	0.739	0.708	0.739	0.716	0.708	0.718	0.719
	Between	0.941	0.943	0.950	0.934	0.940	0.950	0.934	0.944	0.946
MLR	Within	0.914	0.938	0.945	0.928	0.927	0.940	0.925	0.944	0.946
	Between	0.931	0.942	0.955	0.932	0.940	0.941	0.921	0.929	0.939
WLSM	Within	0.604	0.579	0.554	0.541	0.528	0.510	0.531	0.515	0.511
	Between	0.977	0.967	0.957	0.956	0.955	0.954	0.950	0.953	0.953
WLSMV	Within	0.594	0.572	0.562	0.544	0.535	0.520	0.528	0.518	0.500
	Between	0.973	0.968	0.958	0.956	0.954	0.952	0.950	0.951	0.951

Because we are dealing with numerous conditions and data points, a summary table by estimator is helpful to have a clearer picture of what is occurring (especially since a recommendation of sample size is by estimator).

Table 2.16

ML Estimator- Balanced

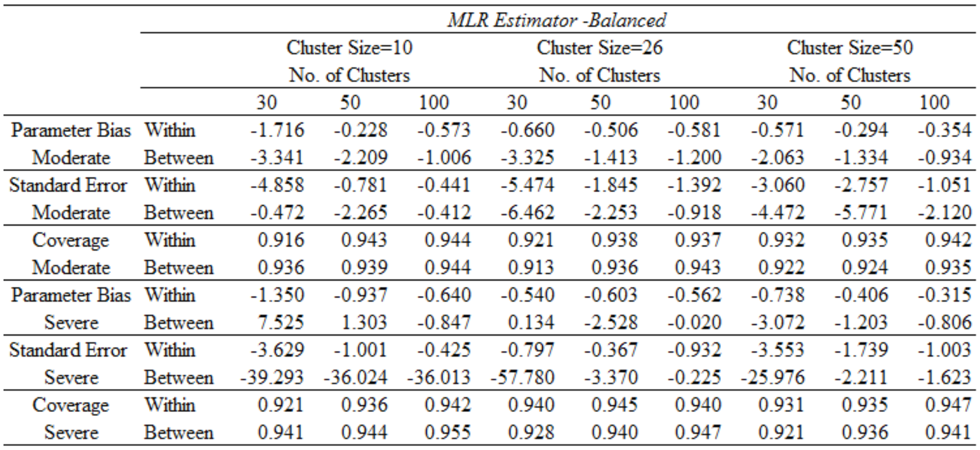


		ML Estimator -Balanced								
		Cluster Size=10			Cluster Size=26			Cluster Size=50		
		No. of Clusters			No. of Clusters			No. of Clusters		
		30	50	100	30	50	100	30	50	100
Parameter Bias	Within	-0.601	-0.664	-0.361	-0.558	-0.630	-0.250	-0.313	-0.483	-0.375
Moderate	Between	-2.681	-2.769	-1.281	-3.300	-1.769	-1.413	-2.591	-1.122	-0.950
Standard Error	Within	-36.899	-35.042	-39.898	-38.007	-36.500	-36.542	-37.671	-35.612	-38.930
Moderate	Between	-13.911	-1.472	0.095	-2.744	-2.206	-2.414	-3.696	-2.227	-1.647
Coverage	Within	0.784	0.797	0.762	0.774	0.786	0.790	0.774	0.785	0.769
Moderate	Between	0.956	0.949	0.951	0.939	0.938	0.940	0.931	0.945	0.949
Parameter Bias	Within	-1.685	-0.519	-0.604	-0.734	-0.287	-0.536	-0.720	-0.733	-0.455
Severe	Between	13.122	0.456	-1.856	1.463	-2.128	-1.556	-2.253	-0.363	-0.781
Standard Error	Within	-43.153	-44.029	-42.152	-44.633	-42.681	-43.724	-43.909	-45.461	-45.075
Severe	Between	-50.068	-27.917	-0.909	-65.845	-3.836	-3.673	-21.651	-1.638	0.080
Coverage	Within	0.737	0.734	0.741	0.733	0.739	0.731	0.720	0.696	0.729
Severe	Between	0.942	0.949	0.961	0.937	0.945	0.941	0.935	0.944	0.952

Table 2.16 provides a clearer picture for the evaluation of the sample size requirement by Muthén and Muthén (2002). If we were solely looking at the three conditions outlined, then ML could not be used at the within level, and one must have at least a sample size of 50 with a small cluster size (ten) at the between level for moderate non-normality. For severe non-normality ML balanced condition, the within level is not recommended with this estimator and a sample size of at least 100 is needed for the small cluster size. At the between level, the required sample size gets smaller as the cluster size gets bigger. Using 10% standard error (SE) bias as a guide, when cluster size is 10, only a sample size of 100 is less than 10% for the severe ML estimator, but once the cluster size increases to 26 then the sample size of 50 is less than the 10% standard error bias.

Table 2.17

MLR Estimator-Balanced



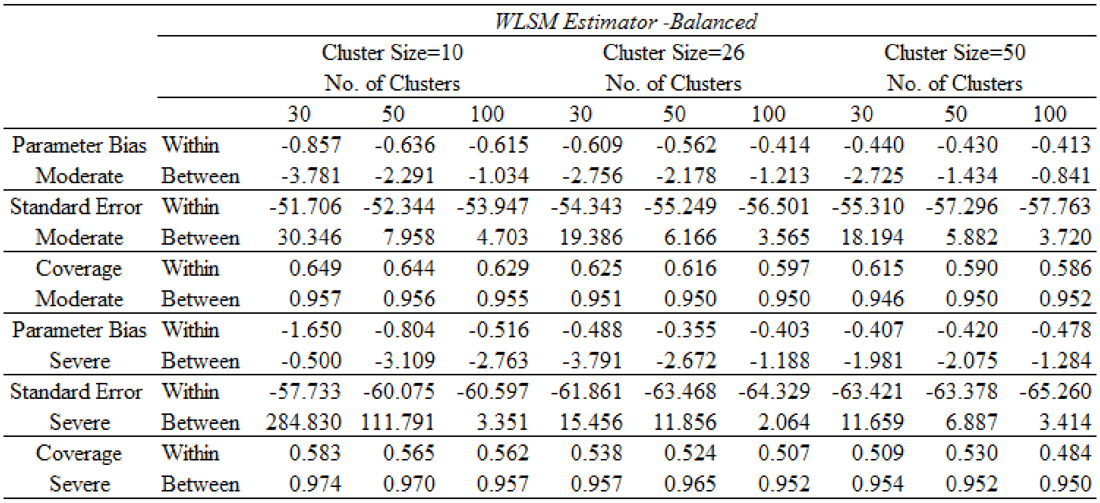
		MLR Estimator -Balanced								
		Cluster Size=10 No. of Clusters			Cluster Size=26 No. of Clusters			Cluster Size=50 No. of Clusters		
		30	50	100	30	50	100	30	50	100
Parameter Bias	Within	-1.716	-0.228	-0.573	-0.660	-0.506	-0.581	-0.571	-0.294	-0.354
Moderate	Between	-3.341	-2.209	-1.006	-3.325	-1.413	-1.200	-2.063	-1.334	-0.934
Standard Error	Within	-4.858	-0.781	-0.441	-5.474	-1.845	-1.392	-3.060	-2.757	-1.051
Moderate	Between	-0.472	-2.265	-0.412	-6.462	-2.253	-0.918	-4.472	-5.771	-2.120
Coverage	Within	0.916	0.943	0.944	0.921	0.938	0.937	0.932	0.935	0.942
Moderate	Between	0.936	0.939	0.944	0.913	0.936	0.943	0.922	0.924	0.935
Parameter Bias	Within	-1.350	-0.937	-0.640	-0.540	-0.603	-0.562	-0.738	-0.406	-0.315
Severe	Between	7.525	1.303	-0.847	0.134	-2.528	-0.020	-3.072	-1.203	-0.806
Standard Error	Within	-3.629	-1.001	-0.425	-0.797	-0.367	-0.932	-3.553	-1.739	-1.003
Severe	Between	-39.293	-36.024	-36.013	-57.780	-3.370	-0.225	-25.976	-2.211	-1.623
Coverage	Within	0.921	0.936	0.942	0.940	0.945	0.940	0.931	0.935	0.947
Severe	Between	0.941	0.944	0.955	0.928	0.940	0.947	0.921	0.936	0.941

Table 2.17 encapsulates the MLR balanced results for the three Muthén and Muthén (2002) criteria (i.e. that the parameter and standard error bias be less than 10% and the coverage be between .91 and .98). For moderate non-normality, a sample size of 30 seems to be okay. However, a sample size of 50 is needed with at least a cluster size of 26 for severe non-normality. If one want to look at between and within SEM data, MLR is the best possible estimator depending on the severity of non-normality.

Consult Table 2.18 for the WLSM estimator. Like the ML estimator, the within level is not recommended. At the between level (i.e. balanced, moderate non-normality) a sample size of at least 50 is recommended and cluster size seems not to matter. For severe non-normality, a sample size of more than 100 is needed for small cluster size. A sample size of 50 will suffice when cluster size is at least 26.

Table 2.18

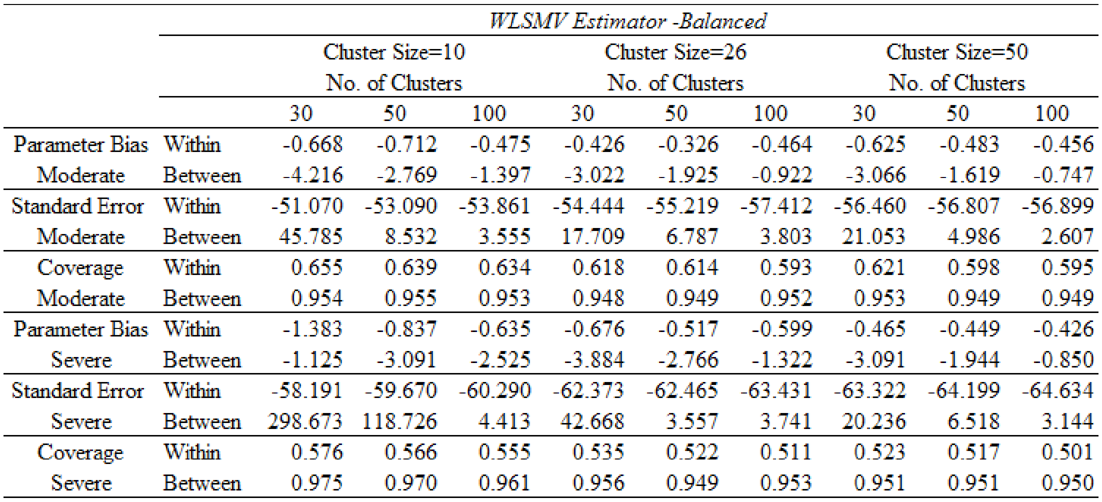
WLSM Estimator-Balanced



		WLSM Estimator -Balanced								
		Cluster Size=10			Cluster Size=26			Cluster Size=50		
		No. of Clusters			No. of Clusters			No. of Clusters		
		30	50	100	30	50	100	30	50	100
Parameter Bias	Within	-0.857	-0.636	-0.615	-0.609	-0.562	-0.414	-0.440	-0.430	-0.413
Moderate	Between	-3.781	-2.291	-1.034	-2.756	-2.178	-1.213	-2.725	-1.434	-0.841
Standard Error	Within	-51.706	-52.344	-53.947	-54.343	-55.249	-56.501	-55.310	-57.296	-57.763
Moderate	Between	30.346	7.958	4.703	19.386	6.166	3.565	18.194	5.882	3.720
Coverage	Within	0.649	0.644	0.629	0.625	0.616	0.597	0.615	0.590	0.586
Moderate	Between	0.957	0.956	0.955	0.951	0.950	0.950	0.946	0.950	0.952
Parameter Bias	Within	-1.650	-0.804	-0.516	-0.488	-0.355	-0.403	-0.407	-0.420	-0.478
Severe	Between	-0.500	-3.109	-2.763	-3.791	-2.672	-1.188	-1.981	-2.075	-1.284
Standard Error	Within	-57.733	-60.075	-60.597	-61.861	-63.468	-64.329	-63.421	-63.378	-65.260
Severe	Between	284.830	111.791	3.351	15.456	11.856	2.064	11.659	6.887	3.414
Coverage	Within	0.583	0.565	0.562	0.538	0.524	0.507	0.509	0.530	0.484
Severe	Between	0.974	0.970	0.957	0.957	0.965	0.952	0.954	0.952	0.950

Table 2.19

WLSMV Estimator-Balanced

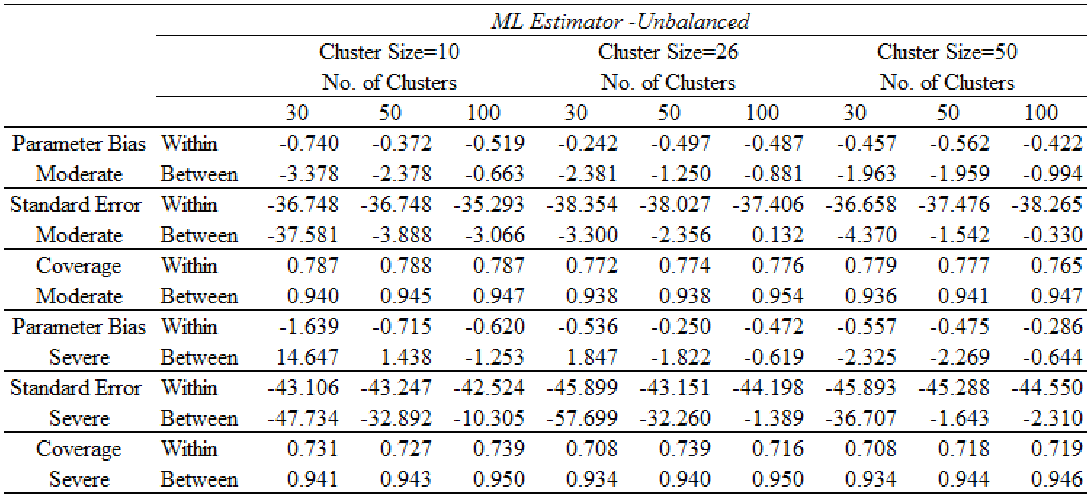


		WLSMV Estimator -Balanced								
		Cluster Size=10			Cluster Size=26			Cluster Size=50		
		No. of Clusters			No. of Clusters			No. of Clusters		
		30	50	100	30	50	100	30	50	100
Parameter Bias	Within	-0.668	-0.712	-0.475	-0.426	-0.326	-0.464	-0.625	-0.483	-0.456
Moderate	Between	-4.216	-2.769	-1.397	-3.022	-1.925	-0.922	-3.066	-1.619	-0.747
Standard Error	Within	-51.070	-53.090	-53.861	-54.444	-55.219	-57.412	-56.460	-56.807	-56.899
Moderate	Between	45.785	8.532	3.555	17.709	6.787	3.803	21.053	4.986	2.607
Coverage	Within	0.655	0.639	0.634	0.618	0.614	0.593	0.621	0.598	0.595
Moderate	Between	0.954	0.955	0.953	0.948	0.949	0.952	0.953	0.949	0.949
Parameter Bias	Within	-1.383	-0.837	-0.635	-0.676	-0.517	-0.599	-0.465	-0.449	-0.426
Severe	Between	-1.125	-3.091	-2.525	-3.884	-2.766	-1.322	-3.091	-1.944	-0.850
Standard Error	Within	-58.191	-59.670	-60.290	-62.373	-62.465	-63.431	-63.322	-64.199	-64.634
Severe	Between	298.673	118.726	4.413	42.668	3.557	3.741	20.236	6.518	3.144
Coverage	Within	0.576	0.566	0.555	0.535	0.522	0.511	0.523	0.517	0.501
Severe	Between	0.975	0.970	0.961	0.956	0.949	0.953	0.951	0.951	0.950

The WLSMV estimator seems to perform well on moderately non-normal data at the between level when the sample size was at least 50. For severe non-normality, the WLSMV sample size recommendation seems to depend on cluster size. When cluster size is low (ten in our case) then a sample of size of 100 is needed. Otherwise, it should be at least 50.

Table 2.20

ML-Unbalanced

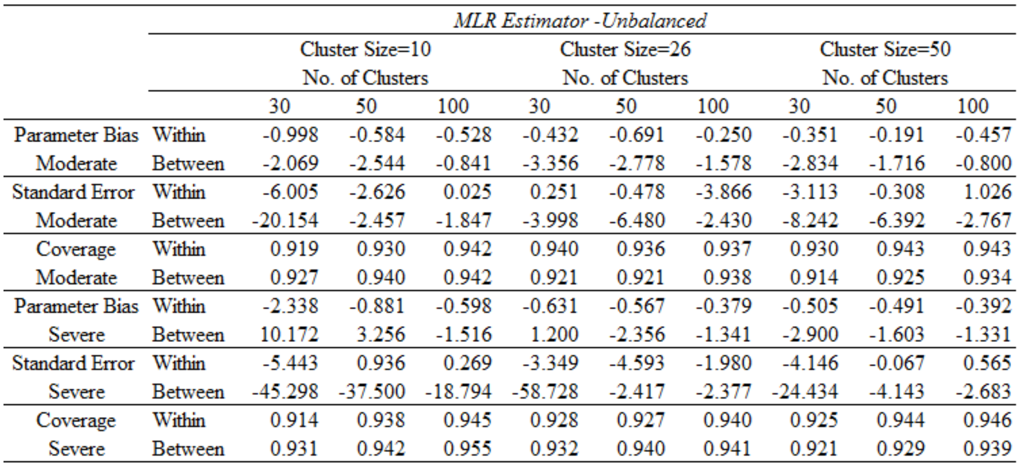


		ML Estimator -Unbalanced								
		Cluster Size=10			Cluster Size=26			Cluster Size=50		
		No. of Clusters			No. of Clusters			No. of Clusters		
		30	50	100	30	50	100	30	50	100
Parameter Bias	Within	-0.740	-0.372	-0.519	-0.242	-0.497	-0.487	-0.457	-0.562	-0.422
Moderate	Between	-3.378	-2.378	-0.663	-2.381	-1.250	-0.881	-1.963	-1.959	-0.994
Standard Error	Within	-36.748	-36.748	-35.293	-38.354	-38.027	-37.406	-36.658	-37.476	-38.265
Moderate	Between	-37.581	-3.888	-3.066	-3.300	-2.356	0.132	-4.370	-1.542	-0.330
Coverage	Within	0.787	0.788	0.787	0.772	0.774	0.776	0.779	0.777	0.765
Moderate	Between	0.940	0.945	0.947	0.938	0.938	0.954	0.936	0.941	0.947
Parameter Bias	Within	-1.639	-0.715	-0.620	-0.536	-0.250	-0.472	-0.557	-0.475	-0.286
Severe	Between	14.647	1.438	-1.253	1.847	-1.822	-0.619	-2.325	-2.269	-0.644
Standard Error	Within	-43.106	-43.247	-42.524	-45.899	-43.151	-44.198	-45.893	-45.288	-44.550
Severe	Between	-47.734	-32.892	-10.305	-57.699	-32.260	-1.389	-36.707	-1.643	-2.310
Coverage	Within	0.731	0.727	0.739	0.708	0.739	0.716	0.708	0.718	0.719
Severe	Between	0.941	0.943	0.950	0.934	0.940	0.950	0.934	0.944	0.946

Similar to the ML-Balanced case (Table 2.16), ML could not be used at the within level. At the between level for moderate normality one must have at least a sample size of 50 with small cluster size (in this case ten). The standard error at the between level is similar for the balanced and non-balanced, moderate non-normality conditions. For the severe non-normality ML unbalanced condition, the within level is not recommended. A sample size of at least 100 is recommended for balanced conditions for the between level. Even when the number of clusters is 100, there is still a large standard error bias. This may or may not be due to sampling variance. Therefore, looking at every piece of information is necessary to get a full picture of what is going on with these estimators. The standard error bias does seems to be different at the between level for the unbalanced versus the balanced non-normal conditions, but again that could be due to sampling variations.

Table 2.21

MLR Unbalanced



		MLR Estimator -Unbalanced								
		Cluster Size=10			Cluster Size=26			Cluster Size=50		
		No. of Clusters			No. of Clusters			No. of Clusters		
		30	50	100	30	50	100	30	50	100
Parameter Bias	Within	-0.998	-0.584	-0.528	-0.432	-0.691	-0.250	-0.351	-0.191	-0.457
Moderate	Between	-2.069	-2.544	-0.841	-3.356	-2.778	-1.578	-2.834	-1.716	-0.800
Standard Error	Within	-6.005	-2.626	0.025	0.251	-0.478	-3.866	-3.113	-0.308	1.026
Moderate	Between	-20.154	-2.457	-1.847	-3.998	-6.480	-2.430	-8.242	-6.392	-2.767
Coverage	Within	0.919	0.930	0.942	0.940	0.936	0.937	0.930	0.943	0.943
Moderate	Between	0.927	0.940	0.942	0.921	0.921	0.938	0.914	0.925	0.934
Parameter Bias	Within	-2.338	-0.881	-0.598	-0.631	-0.567	-0.379	-0.505	-0.491	-0.392
Severe	Between	10.172	3.256	-1.516	1.200	-2.356	-1.341	-2.900	-1.603	-1.331
Standard Error	Within	-5.443	0.936	0.269	-3.349	-4.593	-1.980	-4.146	-0.067	0.565
Severe	Between	-45.298	-37.500	-18.794	-58.728	-2.417	-2.377	-24.434	-4.143	-2.683
Coverage	Within	0.914	0.938	0.945	0.928	0.927	0.940	0.925	0.944	0.946
Severe	Between	0.931	0.942	0.955	0.932	0.940	0.941	0.921	0.929	0.939

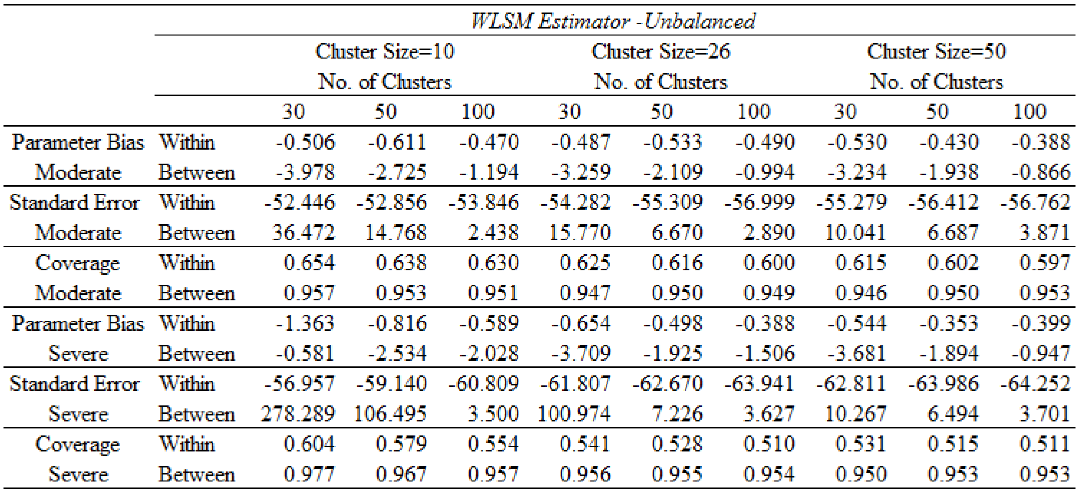
The MLR balanced results were presented in Table 2.17. For unbalanced, moderate non-normality, a sample size of 30 with a corresponding low cluster size has a high standard error bias, unlike the balanced case. Similarly the between level cannot handle low cluster size and low sample size when severe non- normality is present.. Low sample size (30) yielded consistent large negative standard errors under severe non-normality (Table 2.21). Solely looking at the three Muthén and Muthén (2002) criteria, a sample size of more than 100 is recommended under severely unbalanced non-normality when the cluster size is low (10 in our case) and 50 for a moderate cluster size of 26 or more. Although the numbers are somewhat different between the balanced vs. unbalanced case, the conclusions are the same.

For the WLS based estimators, the within level for the unbalanced conditions is not recommended for moderate or severe non-normality. For the between level, the SE bias was high for low cluster size and low sample size for the WLSM estimator under moderate non-normality.

Therefore, a sample size of 100 is needed when cluster size is 10 , and a sample size of 50 is needed when the cluster size is at least 50 (30/26, 15.770 and 50/26, 6.670, Table 2.22). Similarly, under severe non-normality, these sample sizes are within the guideline set by Muthén and Muthén (2002).

Table 2.22

WLSM Unbalanced

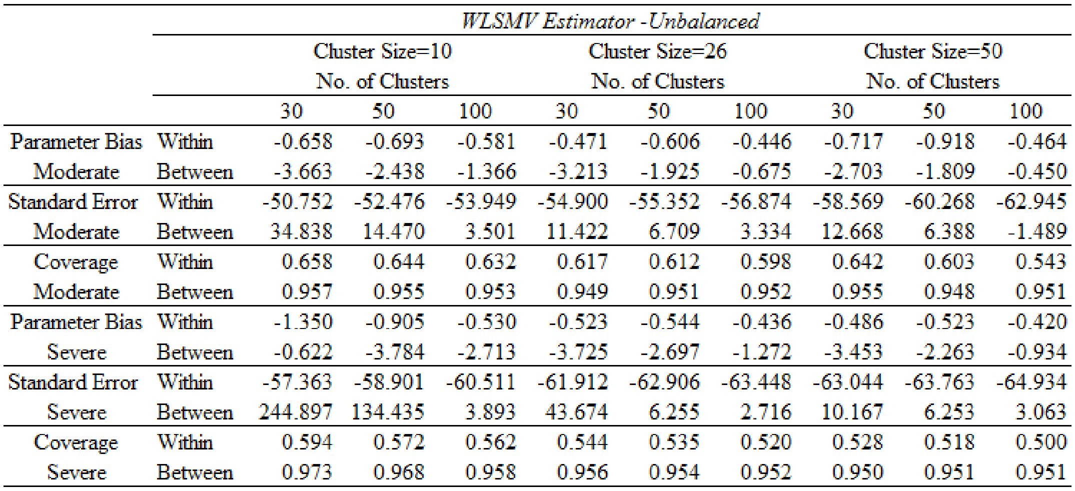


		WLSM Estimator -Unbalanced								
		Cluster Size=10			Cluster Size=26			Cluster Size=50		
		No. of Clusters			No. of Clusters			No. of Clusters		
		30	50	100	30	50	100	30	50	100
Parameter Bias	Within	-0.506	-0.611	-0.470	-0.487	-0.533	-0.490	-0.530	-0.430	-0.388
Moderate	Between	-3.978	-2.725	-1.194	-3.259	-2.109	-0.994	-3.234	-1.938	-0.866
Standard Error	Within	-52.446	-52.856	-53.846	-54.282	-55.309	-56.999	-55.279	-56.412	-56.762
Moderate	Between	36.472	14.768	2.438	15.770	6.670	2.890	10.041	6.687	3.871
Coverage	Within	0.654	0.638	0.630	0.625	0.616	0.600	0.615	0.602	0.597
Moderate	Between	0.957	0.953	0.951	0.947	0.950	0.949	0.946	0.950	0.953
Parameter Bias	Within	-1.363	-0.816	-0.589	-0.654	-0.498	-0.388	-0.544	-0.353	-0.399
Severe	Between	-0.581	-2.534	-2.028	-3.709	-1.925	-1.506	-3.681	-1.894	-0.947
Standard Error	Within	-56.957	-59.140	-60.809	-61.807	-62.670	-63.941	-62.811	-63.986	-64.252
Severe	Between	278.289	106.495	3.500	100.974	7.226	3.627	10.267	6.494	3.701
Coverage	Within	0.604	0.579	0.554	0.541	0.528	0.510	0.531	0.515	0.511
Severe	Between	0.977	0.967	0.957	0.956	0.955	0.954	0.950	0.953	0.953

There is very little difference between the balanced and unbalanced cases for the WLSMV estimator. The conclusions are similar, the between level sample size needs at least 100 when cluster size is low and at least 50 when cluster size is 26 or more (Table 2.23).

Table 2.23

WLSMV Unbalanced



		WLSMV Estimator -Unbalanced								
		Cluster Size=10			Cluster Size=26			Cluster Size=50		
		No. of Clusters			No. of Clusters			No. of Clusters		
		30	50	100	30	50	100	30	50	100
Parameter Bias	Within	-0.658	-0.693	-0.581	-0.471	-0.606	-0.446	-0.717	-0.918	-0.464
Moderate	Between	-3.663	-2.438	-1.366	-3.213	-1.925	-0.675	-2.703	-1.809	-0.450
Standard Error	Within	-50.752	-52.476	-53.949	-54.900	-55.352	-56.874	-58.569	-60.268	-62.945
Moderate	Between	34.838	14.470	3.501	11.422	6.709	3.334	12.668	6.388	-1.489
Coverage	Within	0.658	0.644	0.632	0.617	0.612	0.598	0.642	0.603	0.543
Moderate	Between	0.957	0.955	0.953	0.949	0.951	0.952	0.955	0.948	0.951
Parameter Bias	Within	-1.350	-0.905	-0.530	-0.523	-0.544	-0.436	-0.486	-0.523	-0.420
Severe	Between	-0.622	-3.784	-2.713	-3.725	-2.697	-1.272	-3.453	-2.263	-0.934
Standard Error	Within	-57.363	-58.901	-60.511	-61.912	-62.906	-63.448	-63.044	-63.763	-64.934
Severe	Between	244.897	134.435	3.893	43.674	6.255	2.716	10.167	6.253	3.063
Coverage	Within	0.594	0.572	0.562	0.544	0.535	0.520	0.528	0.518	0.500
Severe	Between	0.973	0.968	0.958	0.956	0.954	0.952	0.950	0.951	0.951

Thus far we have simply looked at the pattern of biases across the various conditions and compared them to the three criteria outlined by Muthén and Muthén (2002). To get a full picture, another criteria was added, the rejection rate for the chi square statistic, which should be no more than 5% (Hoogland and Boosma, 1998). Bradley (1978) suggests this rate could have a more liberal interval of (.025, .075) for chi square being well approximated (as cited in Bandalos, 2006, p. 404). The number of inadmissible solutions and types of errors will also be evaluated for each study, estimator, and condition.

Note, chi-square is a function of the sample size and the Fit function value, so when the sample size is large the chi square will be large and not equal its' approximate degrees of freedom. In addition, chi-square is affected by the degree of non-normality. So for large sample size or non-normality, we expect a fairly large rejection rate.

For study two ML based estimators, there were four types of errors observed after running the Monte Carlo analysis.

Error 1(Miterations):

```
THE MODEL ESTIMATION DID NOT TERMINATE NORMALLY DUE TO A NON-ZERO
DERIVATIVE OF THE OBSERVED-DATA LOGLIKELIHOOD.
THE MCONVERGENCE CRITERION OF THE EM ALGORITHM IS NOT FULFILLED.
CHECK YOUR STARTING VALUES OR INCREASE THE NUMBER OF MITERATIONS.
ESTIMATES CANNOT BE TRUSTED.  THE LOGLIKELIHOOD DERIVATIVE
FOR PARAMETER 18 IS -0.10579199D-01.
```

Error 2 (Residual):

```
WARNING:  THE RESIDUAL COVARIANCE MATRIX (THETA) IS NOT POSITIVE DEFINITE.
THIS COULD INDICATE A NEGATIVE VARIANCE/RESIDUAL VARIANCE FOR AN OBSERVED
VARIABLE, A CORRELATION GREATER OR EQUAL TO ONE BETWEEN TWO OBSERVED
VARIABLES, OR A LINEAR DEPENDENCY AMONG MORE THAN TWO OBSERVED VARIABLES.
CHECK THE RESULTS SECTION FOR MORE INFORMATION.
```

Error 3 (Start Values/Fisher):

```
THE MODEL ESTIMATION DID NOT TERMINATE NORMALLY DUE TO AN ILL-CONDITIONED
 FISHER INFORMATION MATRIX.  CHANGE YOUR MODEL AND/OR STARTING VALUES.
 THE MODEL ESTIMATION DID NOT TERMINATE NORMALLY DUE TO A NON-POSITIVE
 DEFINITE FISHER INFORMATION MATRIX.  THIS MAY BE DUE TO THE STARTING VALUES
 BUT MAY ALSO BE AN INDICATION OF MODEL NONIDENTIFICATION.  THE CONDITION
 NUMBER IS       0.212D-10.
```

Error 4 (Trustworthy Standard Errors):

```
THE STANDARD ERRORS OF THE MODEL PARAMETER ESTIMATES MAY NOT BE
TRUSTWORTHY FOR SOME PARAMETERS DUE TO A NON-POSITIVE DEFINITE
FIRST-ORDER DERIVATIVE PRODUCT MATRIX.  THIS MAY BE DUE TO THE STARTING
VALUES BUT MAY ALSO BE AN INDICATION OF MODEL NONIDENTIFICATION.  THE
CONDITION NUMBER IS       0.918D-12.  PROBLEM INVOLVING PARAMETER 19.
```

The first type of error talks about increasing the number of Miterations allowed in the Expectation Maximization (EM) algorithm. The default number for Mplus is 500 iterations. The EM algorithm is an iterative two step algorithm that reverberates between the expectation step and the maximization step to find the best ML estimates. When presented with a similar message on the statistical model discussion board the advice was to increase the number of Miterations. The error message also mentions Mconvergence which sets the convergence criterion for the EM algorithm. The default Mconvergence value depends on the type of model you run, for type a two level model, .001. I decreased to .00001 due to an error message about saddle points and instructing to decrease its' value. When you increase the number of iterations it takes the program longer to run. By increasing the number of iterations and decreasing the convergence criteria, I did eliminated the saddle point messages and decreased the overall number of Mconvergence errors but I still had a some present and the processing to run the condition increased. It essentially warns you not to trust the estimates given for this particular replication.

The second type of error talks about the Theta matrix. Theta matrix represents the matrix of the residual values. This error says even though the model converged, we have negative residual variance in the replication. Variances should always be positive.

The third type of error suggests changing the starting values. It says the solution you found was not the proper maximum likelihood solution, the solution is unacceptable. Changing the starting value is done in Mplus by changing STARTS. The default starting value is ten random starting values and two optimizations carried out in the final stage. The solution this error would be to increase the number of starts and optimizations. After increasing these values the error message does disappear but one left having to insure the number you received is not a local maximum which is a multistep process that involves rerunning the entire process with the seed generated by the program and the processing time was multifold.

The last or fourth type of error said that the standard errors are not trustworthy for that particular replication. It then instructs you to look at a particular parameter. Parameters 17-20 are from the Theta matrix or Residual matrix. We already know from the second error message that there were problems with the residuals being negative. This particular message occurs when the particular parameter is near its boundary value.

So for this part of the analysis I looked at the percentage of each type of errors and what happens as I increase sample size and/or cluster size. There are no guidelines on what percentage errors are acceptable, so I will use a .05 or less criterion.

Table 2.24

ML Estimator-Balanced Error Analysis

		ML Estimator-Balanced Error Analysis								
		Cluster Size=10			Cluster Size=26			Cluster Size=50		
		No. of Clusters			No. of Clusters			No. of Clusters		
		30	50	100	30	50	100	30	50	100
Degrees of Freedom	DF	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00
Expected Proportion	Expected	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Chi Square	Moderate	4.08	4.14	3.95	4.15	4.17	4.18	4.26	4.09	4.06
	Severe	3.75	3.74	3.80	4.07	4.02	4.13	4.22	4.21	4.05
Observed Proportion	Moderate	0.05	0.05	0.04	0.06	0.07	0.06	0.06	0.06	0.06
	Severe	0.04	0.04	0.04	0.05	0.05	0.06	0.06	0.06	0.05
Successful Reps	Moderate	99.8%	99.9%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	Severe	96.8%	98.6%	100.0%	99.0%	99.9%	100.0%	99.7%	100.0%	100.0%
% Overall Errors	Moderate	17.6%	3.6%	0.1%	3.6%	0.4%	0.0%	2.7%	0.3%	0.0%
	Severe	63.0%	33.4%	5.8%	19.3%	3.1%	1.0%	6.3%	3.0%	0.0%
Miterations	Moderate	0.9%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Severe	33.9%	7.7%	0.0%	1.1%	0.0%	0.0%	0.3%	0.0%	0.0%
Residuals	Moderate	14.7%	3.2%	0.1%	3.6%	0.4%	0.0%	2.7%	0.3%	0.0%
	Severe	24.9%	23.7%	5.3%	15.5%	2.9%	0.1%	5.7%	0.3%	0.0%
Starting Values	Moderate	1.9%	0.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Severe	3.0%	1.5%	0.5%	2.2%	0.2%	0.0%	0.0%	0.0%	0.0%
Standard Errors	Moderate	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Severe	1.2%	0.5%	0.0%	0.5%	0.0%	0.0%	0.3%	0.0%	0.0%

For the ML based estimator, according to Table 2.24, the overall observed portion of chi square that exceeded the critical value was roughly .05, some were as high as .066 (sample size

50/cluster size 26). The proportions of overall errors decreased as sample size increased and as the number of clusters increased. Under moderate non-normality, for the ML estimator, there were large numbers, 14.7 %, of negative residual variances when the cluster size and sample size was low (30/10); so it seems a sample size of 50 and/or cluster size of 26 or more is needed under this condition. When severe non-normality is present a sample size of 100 is need when cluster size is low and at least 50 in general. When the sample size is low, under severe non-normality, 5.3% of the residuals are negative which is close to five percent (30/10).

Table 2.25

ML Estimator-Unbalanced Error Analysis

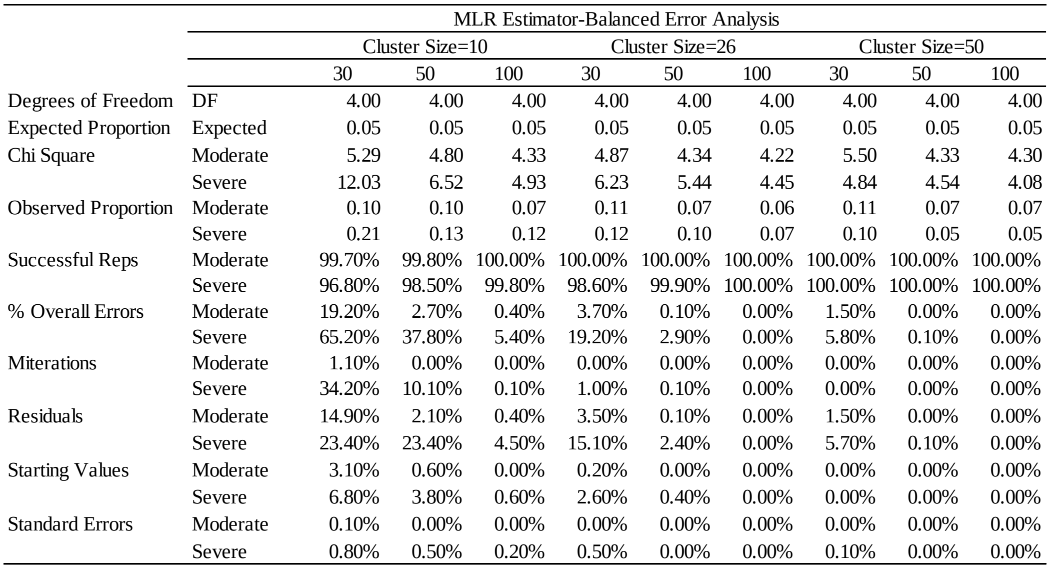
		ML Estimator-Unbalanced Error Analysis								
		Cluster Size=10			Cluster Size=26			Cluster Size=50		
		30	50	100	30	50	100	30	50	100
Degrees of Freedom	DF	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00
Expected Proportion	Expected	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Chi Square	Moderate	4.15	4.08	4.03	4.30	4.11	4.06	4.21	4.02	4.15
	Severe	3.76	3.88	4.00	4.01	4.12	4.01	4.00	3.91	4.15
Observed Proportion	Moderate	0.05	0.05	0.05	0.07	0.05	0.05	0.05	0.06	0.04
	Severe	0.04	0.04	0.05	0.05	0.05	0.06	0.05	0.05	0.07
Successful Reps	Moderate	99.9%	99.9%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	Severe	97.2%	98.1%	99.6%	98.9%	99.8%	100.0%	99.9%	100.0%	100.0%
% Overall Errors	Moderate	18.6%	4.3%	0.2%	4.2%	0.5%	0.0%	2.5%	0.0%	0.0%
	Severe	69.8%	37.9%	6.3%	20.6%	4.2%	0.3%	6.0%	0.6%	0.0%
Miterations	Moderate	1.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Severe	39.6%	8.5%	0.1%	1.4%	0.0%	0.0%	0.0%	0.0%	0.0%
Residuals	Moderate	15.9%	41.0%	2.0%	4.2%	0.5%	0.0%	2.5%	0.0%	0.0%
	Severe	23.7%	23.8%	5.5%	16.8%	3.8%	0.3%	5.7%	0.6%	0.0%
Starting Values	Moderate	1.4%	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Severe	5.0%	5.2%	0.6%	2.2%	0.2%	0.0%	0.2%	0.0%	0.0%
Standard Errors	Moderate	1.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Severe	1.5%	0.4%	0.1%	0.2%	0.2%	0.0%	0.1%	0.0%	0.0%

Table 2.25, shows the error analysis for the unbalanced ML based estimator. The numbers for the unbalanced ML based error analysis are very similar to the balanced ML based error analysis; however, when the sample size is large and cluster size is small under severe non-normality 5.5% of the replications had issues with negative residuals (100/10). My cut off was 5% and this is slightly higher by .2% than the balanced case. A sample size of more than 100 might be

needed for both conditions when the cluster size is more than 100. Other than that, the same pattern of errors, expected proportions, and chi square values exist.

Table 2.26

MLR Estimate-Balanced Error Analysis



		MLR Estimator-Balanced Error Analysis								
		Cluster Size=10			Cluster Size=26			Cluster Size=50		
		30	50	100	30	50	100	30	50	100
Degrees of Freedom	DF	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00
Expected Proportion	Expected	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Chi Square	Moderate	5.29	4.80	4.33	4.87	4.34	4.22	5.50	4.33	4.30
	Severe	12.03	6.52	4.93	6.23	5.44	4.45	4.84	4.54	4.08
Observed Proportion	Moderate	0.10	0.10	0.07	0.11	0.07	0.06	0.11	0.07	0.07
	Severe	0.21	0.13	0.12	0.12	0.10	0.07	0.10	0.05	0.05
Successful Reps	Moderate	99.70%	99.80%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	Severe	96.80%	98.50%	99.80%	98.60%	99.90%	100.00%	100.00%	100.00%	100.00%
% Overall Errors	Moderate	19.20%	2.70%	0.40%	3.70%	0.10%	0.00%	1.50%	0.00%	0.00%
	Severe	65.20%	37.80%	5.40%	19.20%	2.90%	0.00%	5.80%	0.10%	0.00%
Miterations	Moderate	1.10%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	Severe	34.20%	10.10%	0.10%	1.00%	0.10%	0.00%	0.00%	0.00%	0.00%
Residuals	Moderate	14.90%	2.10%	0.40%	3.50%	0.10%	0.00%	1.50%	0.00%	0.00%
	Severe	23.40%	23.40%	4.50%	15.10%	2.40%	0.00%	5.70%	0.10%	0.00%
Starting Values	Moderate	3.10%	0.60%	0.00%	0.20%	0.00%	0.00%	0.00%	0.00%	0.00%
	Severe	6.80%	3.80%	0.60%	2.60%	0.40%	0.00%	0.00%	0.00%	0.00%
Standard Errors	Moderate	0.10%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	Severe	0.80%	0.50%	0.20%	0.50%	0.00%	0.00%	0.10%	0.00%	0.00%

When there is moderate non-normality, for the MLR estimator, a sample size of 50 seems to be needed, when the cluster size is low. A sample size of 30 is needed when there is moderate or large cluster size. Under severe non-normality, when the cluster size is low, a sample size of at least 100 seems to be needed and a sample size of 50 for moderate to large cluster sizes. One should notice that overall there were a large proportions of replications that exceeded the critical chi square value—nowhere close to .05 for most cases, especially, under the severe-normality condition. However, Bradley’s liberal criterion test said that the empirical type one error rate can fall within an acceptable range to be acceptable. It falls between .025 and .075 meeting the Bradley’s liberal criterion (Bradley, 1978) when the nominal alpha is .05 (Hoyle, 2012). This does not quite confirm the theory that chi-square distribution is not well approximated for MLR estimator.

Table 2.27

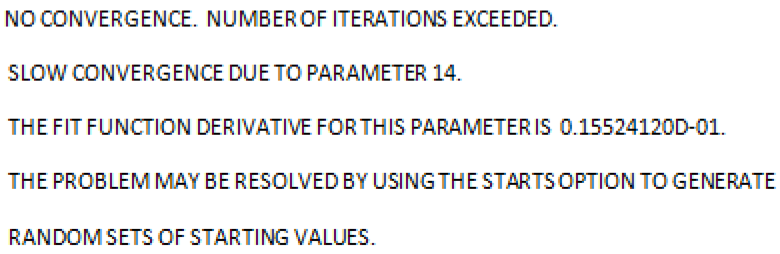
MLR Estimate-Unbalanced Error Analysis

		MLR Estimator-Unbalanced Error Analysis								
		Cluster Size=10			Cluster Size=26			Cluster Size=50		
		30	50	100	30	50	100	30	50	100
Degrees of Freedom	DF	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00
Expected Proportion	Expected	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Chi Square	Moderate	4.85	4.76	4.19	4.95	4.69	4.17	5.20	4.42	4.09
	Severe	11.541	7.102	5.371	5.328	4.757	4.213	4.797	4.513	4.181
Observed Proportion	Moderate	0.09	0.08	0.07	0.10	0.10	0.06	0.11	0.09	0.05
	Severe	0.22	0.13	0.11	0.13	0.10	0.07	0.11	0.09	0.07
Successful Reps	Moderate	99.90%	100.00%	100.00%	99.90%	100.00%	100.00%	100.00%	100.00%	100.00%
	Severe	63.40%	97.80%	99.80%	99.50%	99.90%	100.00%	99.90%	100.00%	100.00%
% Overall Errors	Moderate	21.70%	3.30%	0.30%	4.40%	0.30%	0.10%	1.80%	0.30%	0.00%
	Severe	63.00%	36.00%	8.20%	18.50%	3.40%	0.00%	5.90%	0.40%	0.00%
Miterations	Moderate	1.40%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	Severe	36.70%	8.90%	0.20%	1.30%	0.00%	0.00%	0.00%	0.00%	0.00%
Residuals	Moderate	16.60%	3.30%	0.30%	4.00%	0.30%	0.10%	1.80%	0.30%	0.00%
	Severe	22.50%	21.90%	7.40%	15.30%	3.20%	0.00%	5.80%	0.40%	0.00%
Starting Values	Moderate	3.40%	0.00%	0.00%	0.40%	0.00%	0.00%	0.00%	0.00%	0.00%
	Severe	3.20%	4.70%	0.60%	1.40%	0.20%	0.00%	0.00%	0.00%	0.00%
Standard Errors	Moderate	0.30%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	Severe	0.60%	0.50%	0.00%	0.50%	0.00%	0.00%	0.10%	0.00%	0.00%

When there is moderate non-normality, for the MLR estimator given unbalanced condition, a sample size of 50 seems to be needed for small cluster size and a sample size of 30 for moderate to large cluster sizes. Similarly to the non-balanced case, under severe non-normality, when the cluster size is low, the number of clusters needed seems to be and 50 clusters for moderate to large cluster sizes.

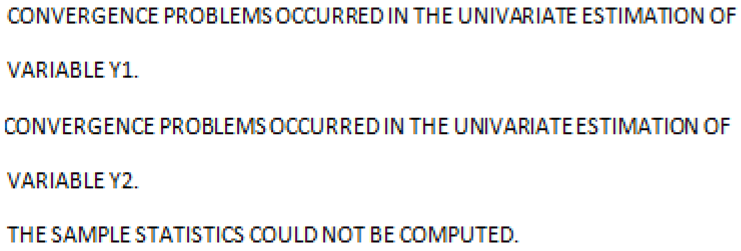
The weighted least squares (WLS) based estimators number of errors were lower than the ML based estimators, and the types of errors found were different from the maximum likelihood (ML) based estimators. There were three errors found when using the WLS based estimators:

Error 5 (Slow Convergence):



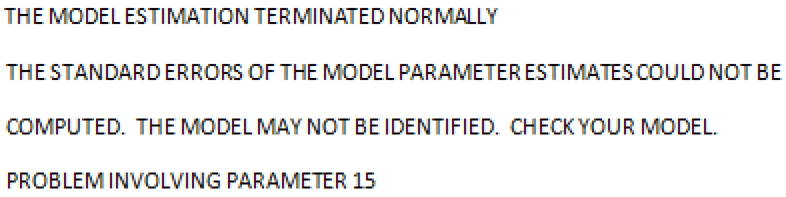
```
NO CONVERGENCE.  NUMBER OF ITERATIONS EXCEEDED.

SLOW CONVERGENCE DUE TO PARAMETER 14.

THE FIT FUNCTION DERIVATIVE FOR THIS PARAMETER IS  0.15524120D-01.

THE PROBLEM MAY BE RESOLVED BY USING THE STARTS OPTION TO GENERATE

RANDOM SETS OF STARTING VALUES.
```

Error 6 (Estimation of Variable):



```
CONVERGENCE PROBLEMS OCCURRED IN THE UNIVARIATE ESTIMATION OF

VARIABLE Y1.

CONVERGENCE PROBLEMS OCCURRED IN THE UNIVARIATE ESTIMATION OF

VARIABLE Y2.

THE SAMPLE STATISTICS COULD NOT BE COMPUTED.
```

Error 7 (Standard Errors):



```
THE MODEL ESTIMATION TERMINATED NORMALLY

THE STANDARD ERRORS OF THE MODEL PARAMETER ESTIMATES COULD NOT BE

COMPUTED.  THE MODEL MAY NOT BE IDENTIFIED.  CHECK YOUR MODEL.

PROBLEM INVOLVING PARAMETER 15
```

Parameters 14 and 15 referred to the Lambda matrix, which is the loading matrix. All these errors referenced difficulty estimating the loadings either it was slow to converge, did not converge for a subset of the indicators (y1-y4), or the standard errors could not be estimated because the loadings were zero.

Table 2.28

WLSM Estimate-Balanced Error Analysis

		WLSM Estimator-Balanced Error Analysis								
		Cluster Size=10			Cluster Size=26			Cluster Size=50		
		30	50	100	30	50	100	30	50	100
Degrees of Freedom	DF	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00
Expected Proportion	Expected	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
Chi Square	Moderate	3.80	4.12	4.22	3.94	4.24	4.40	4.05	4.20	4.33
	Severe	3.58	4.05	4.30	4.07	4.31	4.48	4.18	4.42	4.70
Observed Proportion	Moderate	0.05	0.07	0.07	0.06	0.07	0.08	0.07	0.07	0.07
	Severe	0.05	0.07	0.08	0.07	0.08	0.09	0.08	0.08	0.10
Successful Reps	Moderate	99.70%	100.00%	100.00%	100.00%	100.00%	100.00%	99.90%	100.00%	100.00%
	Severe	91.34%	98.09%	99.90%	99.62%	99..95%	99..95%	99.90%	99.98%	100.00%
% Overall Errors	Moderate	0.28%	0.00%	0.00%	0.02%	0.00%	0.00%	0.07%	0.01%	0.02%
	Severe	9.23%	1.97%	0.10%	0.38%	0.04%	0.06%	0.10%	0.01%	0.00%
Slow Convergence	Moderate	0.04%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%
	Severe	0.86%	0.23%	0.00%	0.06%	0.01%	0.00%	0.00%	0.00%	0.00%
Estimation of Variable	Moderate	0.04%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%	0.01%
	Severe	6.70%	0.76%	0.03%	0.03%	0.00%	0.03%	0.00%	0.00%	0.00%
Standard Errors	Moderate	0.20%	0.00%	0.00%	0.02%	0.00%	0.00%	0.07%	0.00%	0.00%
	Severe	1.67%	0.98%	0.07%	0.29%	0.03%	0.03%	0.10%	0.01%	0.00%

Under moderate non-normality, a sample size of 30 seems reasonable, and at least 50 for severe non-normality. Also, proportion of replications that exceeded the chi square value increased with sample size for the chi-square distribution. Table 2.29, for the unbalanced case, had similar conclusions. Tables 2.30 and 2.31 for the balanced and non-balanced WLSMV estimators, respectively, have similar conclusions for sample size recommendation; however, the observed proportions seem to be slightly less.

Table 2.29

WLSM Estimate-Unbalanced Error Analysis

		WLSM Estimator-Unbalanced Error Analysis								
		Cluster Size=10			Cluster Size=26			Cluster Size=50		
		30	50	100	30	50	100	30	50	100
Degrees of Freedom	DF	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00
Expected Proportion	Expected	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Chi Square	Moderate	3.718	4.064	4.236	3.905	4.172	4.388	3.977	4.083	4.293
	Severe	3.50	3.95	4.37	4.03	4.33	4.45	4.15	4.47	4.63
Observed Proportion	Moderate	0.05	0.065	0.069	0.057	0.071	0.08	0.061	0.064	0.073
	Severe	0.05	0.06	0.08	0.07	0.08	0.08	0.08	0.09	0.10
Successful Reps	Moderate	99.54%	99.98%	100.00%	100.00%	100.00%	100.00%	99.94%	99.97%	99.90%
	Severe	91.17%	98.27%	99.94%	99.50%	99.93%	99.97%	99.86%	99.99%	99.99%
% Overall Errors	Moderate	0.44%	0.01%	0.00%	0.00%	0.00%	0.00%	0.05%	0.01%	0.10%
	Severe	7.98%	1.73%	0.06%	0.05%	0.07%	0.01%	0.04%	0.02%	0.01%
Slow Convergence	Moderate	0.03%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.03%
	Severe	0.09%	0.13%	0.00%	0.08%	0.00%	0.00%	0.00%	0.00%	0.00%
Estimation of Variable	Moderate	0.07%	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%	0.01%	0.01%
	Severe	6.20%	0.93%	0.00%	0.03%	0.00%	0.00%	0.03%	0.00%	0.00%
Standard Errors	Moderate	0.34%	0.01%	0.00%	0.00%	0.00%	0.00%	0.04%	0.01%	0.06%
	Severe	1.69%	0.67%	0.06%	0.39%	0.07%	0.01%	0.01%	0.01%	0.01%

Table 2.30

WLSMV Estimate-Balanced Error Analysis

		WLSMV Estimator-Balanced Error Analysis								
		Cluster Size=10			Cluster Size=26			Cluster Size=50		
		30	50	100	30	50	100	30	50	100
Degrees of Freedom	DF	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00
Expected Proportion	Expected	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Chi Square	Moderate	3.764	4.009	4.241	3.949	4.139	4.34	3.948	4.187	4.27
	Severe	3.60	4.04	4.34	4.03	4.25	4.38	4.07	4.45	4.59
Observed Proportion	Moderate	0.043	0.051	0.062	0.047	0.056	0.065	0.047	0.061	0.062
	Severe	0.04	0.06	0.07	0.05	0.06	0.08	0.06	0.08	0.09
Successful Reps	Moderate	99.56%	99.98%	100.00%	100.00%	100.00%	100.00%	99.94%	99.94%	99.90%
	Severe	91.31%	98.24%	99.90%	99.70%	99.93%	99.92%	99.91%	100.00%	100.00%
% Overall Errors	Moderate	0.49%	0.01%	0.00%	0.00%	0.00%	0.00%	0.05%	0.05%	0.09%
	Severe	9.71%	0.89%	0.10%	0.28%	0.08%	0.09%	0.10%	0.00%	0.00%
Slow Convergence	Moderate	0.03%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%
	Severe	0.86%	0.11%	0.03%	0.06%	0.00%	0.00%	0.01%	0.00%	0.00%
Estimation of Variable	Moderate	0.06%	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%	0.01%	0.01%
	Severe	7.02%	0.09%	0.01%	0.02%	0.00%	0.00%	0.00%	0.00%	0.00%
Standard Errors	Moderate	0.40%	0.01%	0.00%	0.00%	0.00%	0.00%	0.04%	0.04%	0.07%
	Severe	1.83%	0.77%	0.06%	0.20%	0.08%	0.09%	0.09%	0.00%	0.00%

Table 2.31

WLSMV Estimate-Unbalanced Error Analysis

		WLSMV Estimator-Unbalanced Error Analysis								
		Cluster Size=10			Cluster Size=26			Cluster Size=50		
		30	50	100	30	50	100	30	50	100
Degrees of Freedom	DF	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00
Expected Proportion	Expected	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Chi Square	Moderate	3.76	4.02	4.26	3.95	4.22	4.39	3.93	4.16	4.26
	Severe	3.61	4.00	4.33	4.01	4.27	4.48	4.14	4.35	4.57
Observed Proportion	Moderate	0.04	0.50	0.06	0.05	0.06	0.07	0.05	0.06	0.07
	Severe	0.04	0.05	0.07	0.06	0.06	0.08	0.06	0.07	0.08
Successful Reps	Moderate	99.47%	100.00%	100.00%	99.99%	100.00%	100.00%	99.94%	99.90	99.90%
	Severe	90.59%	98.19%	99.92%	99.73%	99.93%	99.87%	99.96%	99.99%	99.99%
% Overall Errors	Moderate	0.53%	0.00%	0.00%	0.01%	0.00%	0.00%	0.06%	0.08%	0.06%
	Severe	10.55%	2.00%	0.08%	0.25%	0.06%	0.13%	0.04%	0.01%	0.01%
Slow Convergence	Moderate	0.06%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	Severe	0.76%	0.24%	0.01%	0.02%	0.01%	0.01%	0.00%	0.00%	0.00%
Estimation of Variable	Moderate	0.06%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%	0.00%
	Severe	8.09%	1.07%	0.03%	0.00%	0.01%	0.03%	0.00%	0.00%	0.00%
Standard Errors	Moderate	0.41%	0.00%	0.00%	0.01%	0.00%	0.00%	0.06%	0.07%	0.06%
	Severe	1.70%	0.69%	0.04%	0.23%	0.04%	0.09%	0.04%	0.01%	0.01%

Results of Study Three

Again, for study three I examined sample size requirement for categorical data. With both studies two and three I did briefly include a very low sample size of size 20. In study two, the very low sample size had similar results as sample size 30 but for study three that was not the case. There was very pronounced differences in sample size 20 versus 30 for study three.

Parameter Bias. Using Muthén and Muthén’s (2002) guideline, the average parameter bias should not exceed 10% for the four estimators, maximum likelihood (ML), robust ML (MLR), weighted least square mean adjusted (WLSM), and weighted least square mean and variance adjusted (WLSMV). There were two sparseness conditions, sparseness I and sparseness II. The data for sparseness I were divided so that 90% of the data was in the first category, then the subsequent categories had 5%, 3%, and 2% of the data. For sparseness II, 50% of the data occurred before the first cutpoint, 30%, and then 10%.

Table 2.32

Average Parameter Bias: Sparseness I

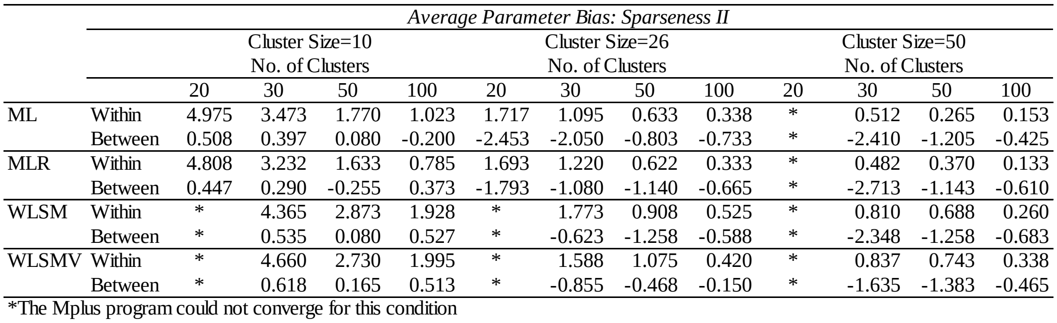
		Average Parameter Bias: Sparseness I											
		Cluster Size=10 No. of Clusters				Cluster Size=26 No. of Clusters				Cluster Size=50 No. of Clusters			
		20	30	50	100	20	30	50	100	20	30	50	100
ML	Within	7.955	4.873	2.393	1.343	2.445	1.548	1.078	0.360	*	0.635	0.538	0.250
	Between	2.850	1.365	0.530	0.340	-0.955	-0.677	-0.063	-0.280	*	-2.003	-0.675	-0.740
MLR	Within	13.635	4.603	2.410	1.250	2.400	1.250	0.878	0.555	*	0.687	0.542	0.243
	Between	7.417	1.833	0.475	0.212	-0.700	-0.530	-0.290	0.083	*	-1.555	-1.225	-0.225
WLSM	Within	10.963	7.078	4.280	2.573	*	2.825	1.445	0.782	*	1.355	0.890	0.355
	Between	2.593	1.307	0.803	0.643	*	-0.553	-0.155	-0.137	*	-1.198	-0.783	0.072
WLSMV	Within	12.090	7.255	4.485	2.670	*	2.178	1.383	0.900	*	1.160	0.910	0.503
	Between	3.388	1.415	1.580	1.138	*	-0.785	-0.530	-0.545	*	-0.818	-0.773	-0.453

*The Mplus program could not converge for this condition

There were no large average parameter bias at the within level except for the MLR, WLSM, WLSMV estimators. The asterisk symbols (*) represent conditions that did not converge. When the sample size was 20 it took so long for the program to run that I halted the program. Because the program ran in a linear line by line order, which ones that did not return a value was arbitrary and due to me halting the program. The Mplus program for those conditions ran for several days and the software did not terminate normally. Based solely on this chart, a sample size of at least 30 is recommended at the between level.

Table 2.33

Average Parameter Bias: Sparseness II



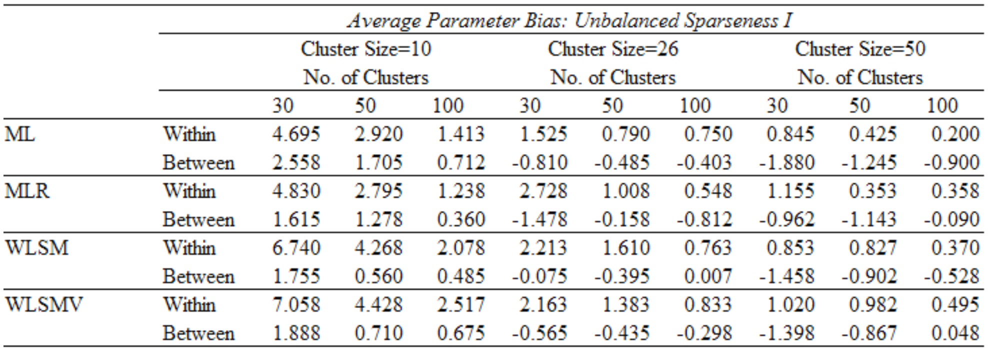
		Average Parameter Bias: Sparseness II											
		Cluster Size=10 No. of Clusters				Cluster Size=26 No. of Clusters				Cluster Size=50 No. of Clusters			
		20	30	50	100	20	30	50	100	20	30	50	100
ML	Within	4.975	3.473	1.770	1.023	1.717	1.095	0.633	0.338	*	0.512	0.265	0.153
	Between	0.508	0.397	0.080	-0.200	-2.453	-2.050	-0.803	-0.733	*	-2.410	-1.205	-0.425
MLR	Within	4.808	3.232	1.633	0.785	1.693	1.220	0.622	0.333	*	0.482	0.370	0.133
	Between	0.447	0.290	-0.255	0.373	-1.793	-1.080	-1.140	-0.665	*	-2.713	-1.143	-0.610
WLSM	Within	*	4.365	2.873	1.928	*	1.773	0.908	0.525	*	0.810	0.688	0.260
	Between	*	0.535	0.080	0.527	*	-0.623	-1.258	-0.588	*	-2.348	-1.258	-0.683
WLSMV	Within	*	4.660	2.730	1.995	*	1.588	1.075	0.420	*	0.837	0.743	0.338
	Between	*	0.618	0.165	0.513	*	-0.855	-0.468	-0.150	*	-1.635	-1.383	-0.465

*The Mplus program could not converge for this condition

There were no large parameter biases for the sparseness II condition. Based solely on this chart and the Muthén and Muthén (2002) criterion, a sample size of at least 30 is recommended.

Table 2.34

Average Parameter Bias: Unbalanced Sparseness I

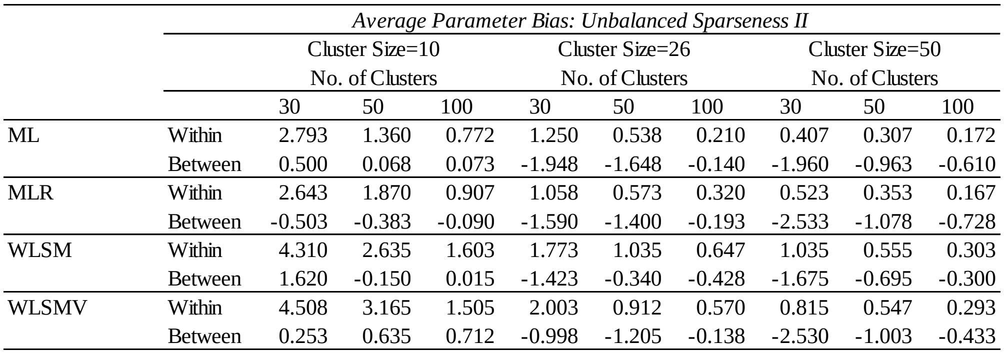


		Average Parameter Bias: Unbalanced Sparseness I								
		Cluster Size=10 No. of Clusters			Cluster Size=26 No. of Clusters			Cluster Size=50 No. of Clusters		
		30	50	100	30	50	100	30	50	100
ML	Within	4.695	2.920	1.413	1.525	0.790	0.750	0.845	0.425	0.200
	Between	2.558	1.705	0.712	-0.810	-0.485	-0.403	-1.880	-1.245	-0.900
MLR	Within	4.830	2.795	1.238	2.728	1.008	0.548	1.155	0.353	0.358
	Between	1.615	1.278	0.360	-1.478	-0.158	-0.812	-0.962	-1.143	-0.090
WLSM	Within	6.740	4.268	2.078	2.213	1.610	0.763	0.853	0.827	0.370
	Between	1.755	0.560	0.485	-0.075	-0.395	0.007	-1.458	-0.902	-0.528
WLSMV	Within	7.058	4.428	2.517	2.163	1.383	0.833	1.020	0.982	0.495
	Between	1.888	0.710	0.675	-0.565	-0.435	-0.298	-1.398	-0.867	0.048

Sample size 20 was not included for the unbalanced conditions because of the length of time and lack of convergence. As we see from table 2.34, there was no large parameter bias. Although when the cluster size and sample size is low the within level bias is somewhat large but not over the criterion establish by Muthén and Muthén (2002). A sample size of at least 30 is recommended.

Table 2.35

Average Parameter Bias: Unbalanced Sparseness II



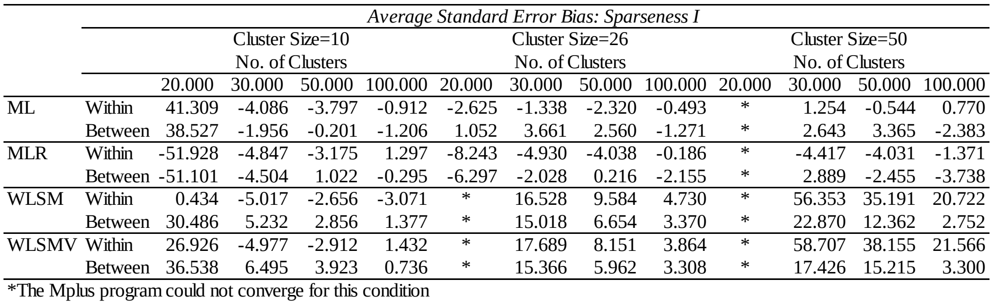
		Average Parameter Bias: Unbalanced Sparseness II								
		Cluster Size=10			Cluster Size=26			Cluster Size=50		
		No. of Clusters			No. of Clusters			No. of Clusters		
		30	50	100	30	50	100	30	50	100
ML	Within	2.793	1.360	0.772	1.250	0.538	0.210	0.407	0.307	0.172
	Between	0.500	0.068	0.073	-1.948	-1.648	-0.140	-1.960	-0.963	-0.610
MLR	Within	2.643	1.870	0.907	1.058	0.573	0.320	0.523	0.353	0.167
	Between	-0.503	-0.383	-0.090	-1.590	-1.400	-0.193	-2.533	-1.078	-0.728
WLSM	Within	4.310	2.635	1.603	1.773	1.035	0.647	1.035	0.555	0.303
	Between	1.620	-0.150	0.015	-1.423	-0.340	-0.428	-1.675	-0.695	-0.300
WLSMV	Within	4.508	3.165	1.505	2.003	0.912	0.570	0.815	0.547	0.293
	Between	0.253	0.635	0.712	-0.998	-1.205	-0.138	-2.530	-1.003	-0.433

Similarly, the unbalanced sparseness II condition did not have problems with large parameter biases. A sample size of at least 30 is recommended.

Standard Error Bias. The next bias consideration is the standard error bias. As a reminder, Muthén and Muthén (2002) set a criterion that states the standard error bias should not exceed 10%. Included in this study are four estimators are maximum likelihood (ML), robust ML (MLR), weighted least square mean adjusted (WLSM), and weighted least square mean and variance adjusted (WLSMV); and two sparseness conditions I and II (with condition I being more sparse).

Table 2.36

Average Standard Error Bias: Sparseness I



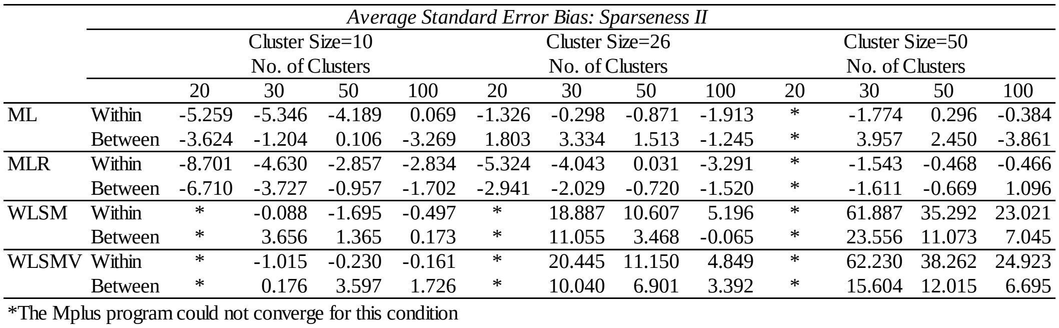
		Average Standard Error Bias: Sparseness I											
		Cluster Size=10 No. of Clusters				Cluster Size=26 No. of Clusters				Cluster Size=50 No. of Clusters			
		20.000	30.000	50.000	100.000	20.000	30.000	50.000	100.000	20.000	30.000	50.000	100.000
ML	Within	41.309	-4.086	-3.797	-0.912	-2.625	-1.338	-2.320	-0.493	*	1.254	-0.544	0.770
	Between	38.527	-1.956	-0.201	-1.206	1.052	3.661	2.560	-1.271	*	2.643	3.365	-2.383
MLR	Within	-51.928	-4.847	-3.175	1.297	-8.243	-4.930	-4.038	-0.186	*	-4.417	-4.031	-1.371
	Between	-51.101	-4.504	1.022	-0.295	-6.297	-2.028	0.216	-2.155	*	2.889	-2.455	-3.738
WLSM	Within	0.434	-5.017	-2.656	-3.071	*	16.528	9.584	4.730	*	56.353	35.191	20.722
	Between	30.486	5.232	2.856	1.377	*	15.018	6.654	3.370	*	22.870	12.362	2.752
WLSMV	Within	26.926	-4.977	-2.912	1.432	*	17.689	8.151	3.864	*	58.707	38.155	21.566
	Between	36.538	6.495	3.923	0.736	*	15.366	5.962	3.308	*	17.426	15.215	3.300

*The Mplus program could not converge for this condition

There were large between standard errors when the sample size 20 and cluster size was low. The within level standard error (SE) bias for sample size 20 and cluster size 10 was extremely low possibly because there were a lot of non-convergent conditions. Also, just as off putting, the SE bias seem to be get worse as cluster size increase for the WLS estimators. This pattern emerged that was not expected. For example, on table 2.36, 30(10) between SE bias was 5.232 for the WLSM estimator but for condition 30(26) 15.018. It seems that for WLSM based estimators that no sample sized can be recommended.

Table 2.37

Average Standard Error Bias: Sparseness II



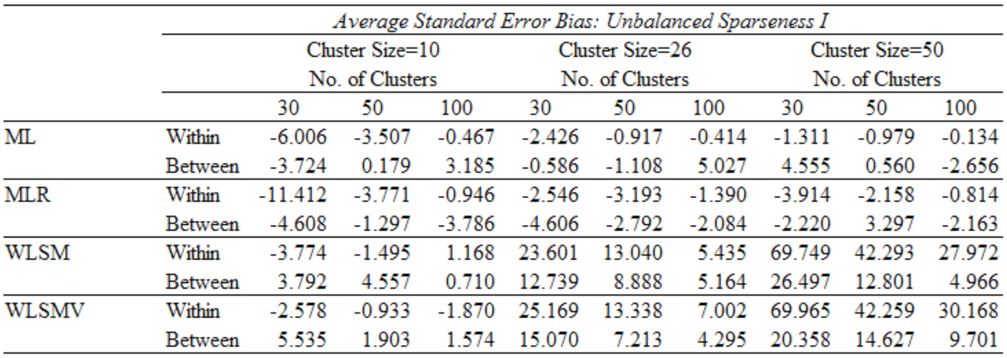
		Average Standard Error Bias: Sparseness II											
		Cluster Size=10 No. of Clusters				Cluster Size=26 No. of Clusters				Cluster Size=50 No. of Clusters			
		20	30	50	100	20	30	50	100	20	30	50	100
ML	Within	-5.259	-5.346	-4.189	0.069	-1.326	-0.298	-0.871	-1.913	*	-1.774	0.296	-0.384
	Between	-3.624	-1.204	0.106	-3.269	1.803	3.334	1.513	-1.245	*	3.957	2.450	-3.861
MLR	Within	-8.701	-4.630	-2.857	-2.834	-5.324	-4.043	0.031	-3.291	*	-1.543	-0.468	-0.466
	Between	-6.710	-3.727	-0.957	-1.702	-2.941	-2.029	-0.720	-1.520	*	-1.611	-0.669	1.096
WLSM	Within	*	-0.088	-1.695	-0.497	*	18.887	10.607	5.196	*	61.887	35.292	23.021
	Between	*	3.656	1.365	0.173	*	11.055	3.468	-0.065	*	23.556	11.073	7.045
WLSMV	Within	*	-1.015	-0.230	-0.161	*	20.445	11.150	4.849	*	62.230	38.262	24.923
	Between	*	0.176	3.597	1.726	*	10.040	6.901	3.392	*	15.604	12.015	6.695

*The Mplus program could not converge for this condition

Similar to sparseness I, it appears for the WLS based estimators the SE bias increases with cluster size so WLS based estimators are not recommended. There is no abnormally large standard error bias for Sparseness I when the sample size is twenty but because of the difficult converging for sample size 20 (cluster size 50) I would recommend a sample size of at least 30 if we are solely looking at this output.

Table 2.38

Average Standard Error Bias: Unbalanced Sparseness I



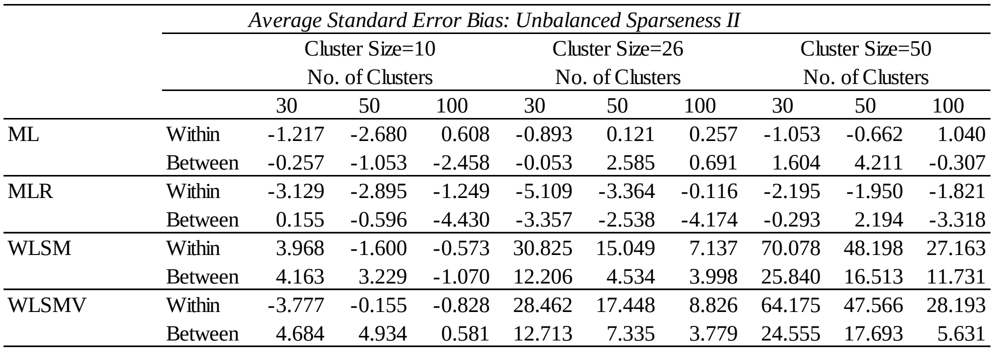
Average Standard Error Bias: Unbalanced Sparseness I

		Cluster Size=10			Cluster Size=26			Cluster Size=50		
		No. of Clusters			No. of Clusters			No. of Clusters		
		30	50	100	30	50	100	30	50	100
ML	Within	-6.006	-3.507	-0.467	-2.426	-0.917	-0.414	-1.311	-0.979	-0.134
	Between	-3.724	0.179	3.185	-0.586	-1.108	5.027	4.555	0.560	-2.656
MLR	Within	-11.412	-3.771	-0.946	-2.546	-3.193	-1.390	-3.914	-2.158	-0.814
	Between	-4.608	-1.297	-3.786	-4.606	-2.792	-2.084	-2.220	3.297	-2.163
WLSM	Within	-3.774	-1.495	1.168	23.601	13.040	5.435	69.749	42.293	27.972
	Between	3.792	4.557	0.710	12.739	8.888	5.164	26.497	12.801	4.966
WLSMV	Within	-2.578	-0.933	-1.870	25.169	13.338	7.002	69.965	42.259	30.168
	Between	5.535	1.903	1.574	15.070	7.213	4.295	20.358	14.627	9.701

The WLS based estimators should be avoided. The ML based estimators are fine except for the within cluster size 10 and sample size 30. A recommended sample size at the between level is 30.

Table 2.39

Average Standard Error Bias: Unbalanced Sparseness II



Average Standard Error Bias: Unbalanced Sparseness II

		Cluster Size=10			Cluster Size=26			Cluster Size=50		
		No. of Clusters			No. of Clusters			No. of Clusters		
		30	50	100	30	50	100	30	50	100
ML	Within	-1.217	-2.680	0.608	-0.893	0.121	0.257	-1.053	-0.662	1.040
	Between	-0.257	-1.053	-2.458	-0.053	2.585	0.691	1.604	4.211	-0.307
MLR	Within	-3.129	-2.895	-1.249	-5.109	-3.364	-0.116	-2.195	-1.950	-1.821
	Between	0.155	-0.596	-4.430	-3.357	-2.538	-4.174	-0.293	2.194	-3.318
WLSM	Within	3.968	-1.600	-0.573	30.825	15.049	7.137	70.078	48.198	27.163
	Between	4.163	3.229	-1.070	12.206	4.534	3.998	25.840	16.513	11.731
WLSMV	Within	-3.777	-0.155	-0.828	28.462	17.448	8.826	64.175	47.566	28.193
	Between	4.684	4.934	0.581	12.713	7.335	3.779	24.555	17.693	5.631

It appears for ML based estimators a sample size of 30 is fine, for unbalanced with sparseness II. The WLS based estimators are not recommended for this analysis.

Coverage. Muthén and Muthén (2002) had for their last requirement that coverage be within the interval (.91, .98) for assessing sample size requirements in structural equation modeling.

Table 2.40

Average Coverage: Sparseness I

		Average Coverage: Sparseness I											
		Cluster Size=10				Cluster Size=26				Cluster Size=50			
		No. of Clusters				No. of Clusters				No. of Clusters			
		20	30	50	100	20	30	50	100	20	30	50	100
ML	Within	0.960	0.962	0.951	0.956	0.958	0.956	0.954	0.952	*	0.957	0.952	0.955
	Between	0.935	0.939	0.946	0.945	0.930	0.939	0.952	0.945	*	0.927	0.948	0.935
MLR	Within	0.947	0.955	0.949	0.954	0.930	0.936	0.937	0.952	*	0.931	0.937	0.946
	Between	0.912	0.932	0.942	0.946	0.907	0.919	0.935	0.932	*	0.924	0.926	0.934
WLSM	Within	0.960	0.962	0.961	0.951	*	0.971	0.961	0.956	*	0.979	0.977	0.969
	Between	0.950	0.947	0.951	0.949	*	0.949	0.951	0.948	*	0.951	0.951	0.945
WLSMV	Within	0.956	0.957	0.958	0.961	*	0.969	0.959	0.957	*	0.983	0.976	0.973
	Between	0.945	0.949	0.954	0.953	*	0.947	0.942	0.945	*	0.941	0.959	0.944

*The Mplus program could not converge for this condition

Like study two, coverage does not appear to be problematic at the between level or at the between level.

Table 2.41

Average Coverage: Sparseness II

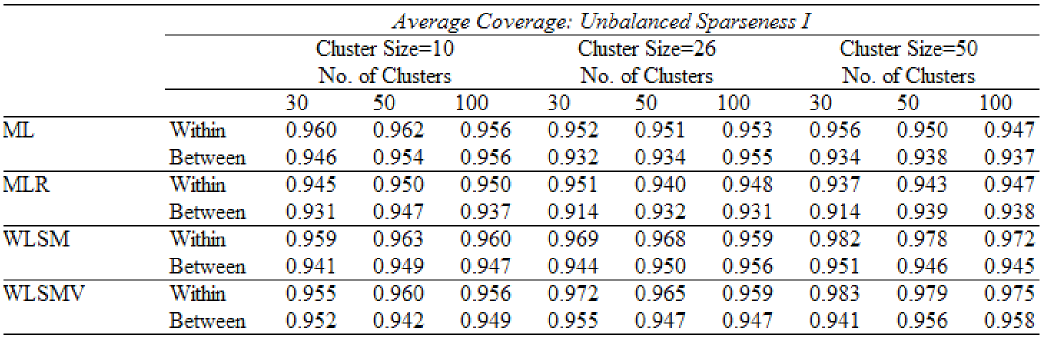
		Average Coverage: Sparseness II											
		Cluster Size=10				Cluster Size=26				Cluster Size=50			
		No. of Clusters				No. of Clusters				No. of Clusters			
		20	30	50	100	20	30	50	100	20	30	50	100
ML	Within	0.962	0.954	0.947	0.955	0.951	0.957	0.951	0.948	*	0.945	0.951	0.952
	Between	0.929	0.941	0.949	0.935	0.939	0.935	0.938	0.936	*	0.922	0.938	0.938
MLR	Within	0.941	0.942	0.946	0.944	0.934	0.935	0.948	0.941	*	0.940	0.945	0.949
	Between	0.915	0.925	0.931	0.939	0.906	0.916	0.925	0.936	*	0.905	0.925	0.941
WLSM	Within	*	0.961	0.954	0.954	*	0.969	0.964	0.955	*	0.982	0.973	0.969
	Between	*	0.947	0.947	0.947	*	0.949	0.936	0.942	*	0.949	0.949	0.953
WLSMV	Within	*	0.958	0.961	0.955	*	0.970	0.966	0.956	*	0.985	0.980	0.973
	Between	*	0.946	0.953	0.958	*	0.940	0.951	0.943	*	0.951	0.949	0.944

*The Mplus program could not converge for this condition

For those conditions that converged, the average coverage was at or above .91.

Table 2.42

Average Coverage: Unbalanced Sparseness I

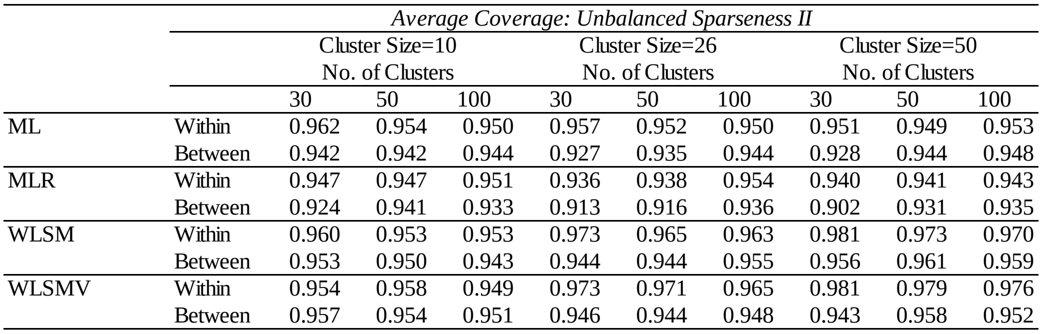


		Average Coverage: Unbalanced Sparseness I								
		Cluster Size=10 No. of Clusters			Cluster Size=26 No. of Clusters			Cluster Size=50 No. of Clusters		
		30	50	100	30	50	100	30	50	100
ML	Within	0.960	0.962	0.956	0.952	0.951	0.953	0.956	0.950	0.947
	Between	0.946	0.954	0.956	0.932	0.934	0.955	0.934	0.938	0.937
MLR	Within	0.945	0.950	0.950	0.951	0.940	0.948	0.937	0.943	0.947
	Between	0.931	0.947	0.937	0.914	0.932	0.931	0.914	0.939	0.938
WLSM	Within	0.959	0.963	0.960	0.969	0.968	0.959	0.982	0.978	0.972
	Between	0.941	0.949	0.947	0.944	0.950	0.956	0.951	0.946	0.945
WLSMV	Within	0.955	0.960	0.956	0.972	0.965	0.959	0.983	0.979	0.975
	Between	0.952	0.942	0.949	0.955	0.947	0.947	0.941	0.956	0.958

Coverage was at or above .91 for all conditions.

Table 2.43

Average Coverage: Unbalanced Sparseness II



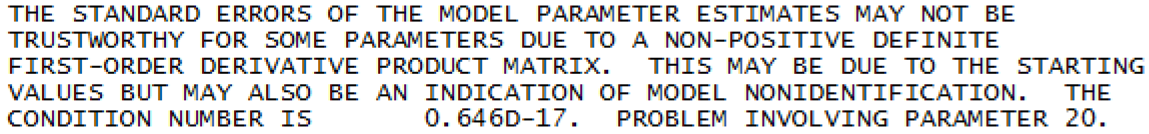
		Average Coverage: Unbalanced Sparseness II								
		Cluster Size=10 No. of Clusters			Cluster Size=26 No. of Clusters			Cluster Size=50 No. of Clusters		
		30	50	100	30	50	100	30	50	100
ML	Within	0.962	0.954	0.950	0.957	0.952	0.950	0.951	0.949	0.953
	Between	0.942	0.942	0.944	0.927	0.935	0.944	0.928	0.944	0.948
MLR	Within	0.947	0.947	0.951	0.936	0.938	0.954	0.940	0.941	0.943
	Between	0.924	0.941	0.933	0.913	0.916	0.936	0.902	0.931	0.935
WLSM	Within	0.960	0.953	0.953	0.973	0.965	0.963	0.981	0.973	0.970
	Between	0.953	0.950	0.943	0.944	0.944	0.955	0.956	0.961	0.959
WLSMV	Within	0.954	0.958	0.949	0.973	0.971	0.965	0.981	0.979	0.976
	Between	0.957	0.954	0.951	0.946	0.944	0.948	0.943	0.958	0.952

The average coverage was fine for all conditions except one MLR 30(50) was .902 (Table 2.40-2.43). Coverage was always usually lower when sample size was lowest.

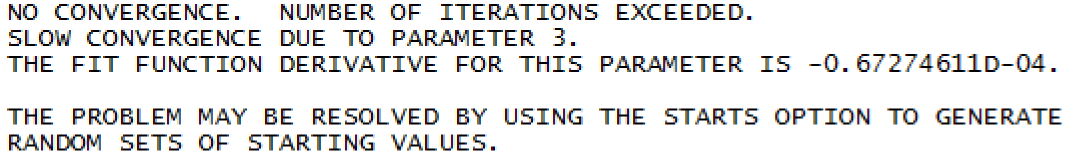
Putting all the information together there seems to be a general recommendation of a sample size of 30 and cluster size of 10 for ML based estimators but to ensure the absolute best coverage a sample size of 50 with at least 10 clusters should be fine for MLR. There were no sample size recommendations for WLS based estimators.

The next step of the process was to analyze the errors and look at the chi-square. However, for categorical variables and maximum likelihood estimators, the chi square fit statistics is not available (Mplus Discussion Board, 2013).   Therefore, for study three we cannot include the chi square fit statistics. Analyzing only the errors for study three was a simple process. Mplus did not print any errors and ran all replications when the sample size was at least 30 for all estimators. The only time errors were printed or encountered was when sample size was 20. Note, at that low sample size, as I may have stated before, not only did the program take over a month to run it was difficult to converge as sample size and cluster sized increased for all estimators.

There were two errors displayed:



```
THE STANDARD ERRORS OF THE MODEL PARAMETER ESTIMATES MAY NOT BE
TRUSTWORTHY FOR SOME PARAMETERS DUE TO A NON-POSITIVE DEFINITE
FIRST-ORDER DERIVATIVE PRODUCT MATRIX.  THIS MAY BE DUE TO THE STARTING
VALUES BUT MAY ALSO BE AN INDICATION OF MODEL NONIDENTIFICATION.  THE
CONDITION NUMBER IS       0.646D-17.  PROBLEM INVOLVING PARAMETER 20.
```



```
NO CONVERGENCE.  NUMBER OF ITERATIONS EXCEEDED.
SLOW CONVERGENCE DUE TO PARAMETER 3.
THE FIT FUNCTION DERIVATIVE FOR THIS PARAMETER IS -0.67274611D-04.

THE PROBLEM MAY BE RESOLVED BY USING THE STARTS OPTION TO GENERATE
RANDOM SETS OF STARTING VALUES.
```

The primary error message was a warning about the standard errors of the model not being trustworthy. Only one estimator, WLSM, had a slow convergence error and this error was specified only once.

The standard errors not being estimated pointed to problem with the tau matrix at the between level which were the intercepts. Upon further investigation the intercepts seemed to be identified but there was not enough information in the output pointing to what was causing the error. It could be that the matrix was not positive definite.   What should be noted is that this error, all errors, went away as soon as the sample size increased. Also, errors decreased as cluster size increased and for less sparse data, see Table 2.44 below. For the WLS based estimated very limited to no errors were reported.

Based on the error analysis, one should not use sample size twenty because it fully studied and implemented in this study and the very few conditions converged.

Table 2.44

Sample Size 20 Error Analysis

		Sample Size 20 Error Analysis											
		Cluster Size=10				Cluster Size=26				Cluster Size=50			
		ML	MLR	WLSM	WLSMV	ML	MLR	WLSM	WLSMV	ML	MLR	WLSM	WLSMV
Successful Reps	Sparseness I	99.95%	99.80%	100.00%	100.00%	100.00%	100.00%	*	*	*	*	*	*
	Sparseness II	100.00%	100.00%	*	*	100.00%	100.00%	*	*	*	*	*	*
% Overall Errors	Sparseness I	56.05%	57.35%	0.50%	0.40%	12.30%	13.15%	*	*	*	*	*	*
	Sparseness II	4.95%	3.95%	*	*	0.50%	0.55%	*	*	*	*	*	*
Standard Errors	Sparseness I	56.05%	57.05%	0.45%	0.40%	12.30%	13.15%	*	*	*	*	*	*
	Sparseness II	4.95%	3.95%	*	*	0.50%	0.55%	*	*	*	*	*	*
Slow Converge	Sparseness I	0.00%	0.00%	0.05%	0.00%	0.00%	0.00%	*	*	*	*	*	*
	Sparseness II	0.00%	0.00%	*	*	0.00%	0.00%	*	*	*	*	*	*

CHAPTER FIVE

STUDY TWO CONCLUSION

For Monte Carlo study two, I had three questions I wanted to answer when I began my study: (a) Does sample size requirement for non-normal continuous data depend on the estimation method?; (b) Is the sample size requirement greater for normal or non-normal continuous data for the respective estimation method?; and (c) Does the presence or absence of unbalanced clusters affect the sample size requirement for non-normal continuous data?

In chapter four, I looked at numerous conditions, used Muthén and Muthén (2002) criteria, and looked at the error rate to determine the best sample size. Not all the conclusions were the same, so putting all the information together in a simplified way to obtain a more accurate picture was necessary. Table 2.45 shows all the recommendations given in chapter four by sample size (cluster size) and for each estimator.

Table 2.45

Between level Minimum Sample (Cluster Size)-Balanced

		Between Level Minimum Sample Size (Cluster Size) Recommendation			
		Parameter Error Bias	Standard Error Bias	95% Coverage	Error Analysis
ML	Moderate	*	50(10)	*	50(10)
	Severe	50(10)	100(10); 50(26);50(50)	*	100(10)?;50(26);50(50)
MLR	Moderate	*	*	*	50(10)
	Severe	*	50(26);50(50)	*	100(10)?;50(26);50(50)
WLSM	Moderate	*	50(10);50(26);50(50)	*	30(10)
	Severe	*	100(10);100(26);50(50)	*	50(10)
WLSMV	Moderate	*	50(10);50(26);50(50)	*	30(10)
	Severe	*	100(10);50(26);50(50)	*	50(10)

* No large bias in all conditions.

There was no large parameter bias present for the ML moderate non-normality condition but under severe non-normality there was a large parameter bias present when for the sample size lower than 50 (so sample size 30 and cluster size 10; 30(10) had large parameter bias), according to the Table 2.45. Looking at the ML estimator, under moderate non-normality, a sample size of 50 and cluster size of 10 is all that is needed (if you want a lower sample size then a larger cluster size is needed 30 (26) was also acceptable). Note, this is the between level sample size requirement. The within level does not perform very well with these estimators unless you are using the MLR based estimator (as seen from this study). This sample size requirement for the between and within using MLR depends on the severity of the non-normality. Although researchers might want to study the between and within levels, for this study I am simply looking at the between level as previous researcher such as Hox and others have done. Under severe non-normality, the ML based estimator requires a large sample size, at least 100 when the cluster size is small, but in general a sample size of 50 will suffice if you have a moderate cluster size of 26 or more. The robust Maximum Likelihood (MLR) estimator between-level had similar finding with a sample size of 50 and cluster size of 10 being the minimum needed under moderate non-normality and for severe non-normality a sample size of 50 with cluster size of 26 or more is needed (with a large sample size of 100 or more being needed for a small cluster size of 10).

The WLS estimators had similar recommendations under moderate non-normality, 50 (10), sample size (cluster size); however, under severe non-normality these recommendations diverge from the ML based estimations. For WLSM, if the cluster size is small or moderate, a sample size of 100 is needed, but when the cluster is as large as 50 a lower sample size of 50 is needed. For WLSMV, a sample size of 100 is needed only for small cluster sizes and a sample size of 50 is needed for moderate cluster sizes.

The severity of the non-normality seems to affect the sample sizes and also the estimator seems to affect the sample size recommendation when severe non-normality is present. Table 2.46, below, shows the same recommendation but only for the unbalanced condition.

Table 2.46

Between level Minimum Sample (Cluster Size)-Unbalanced

		Unbalanced Between Level Minimum Sample Size (Cluster Size) Recommendation			
		Parameter Error Bias	Standard Error Bias	95% Coverage	Error Analysis
ML	Moderate	*	50(10)	*	50(10)
	Severe	50(10)	100(26);50(50)	*	100(10)?;50(26);50(50)
MLR	Moderate	*	50(10)	*	50(10)
	Severe	50(10)	50(26);50(50)	*	50(26);50(50)
WLSM	Moderate	*	100(10);50(26);50(50)	*	30(10)
	Severe	*	100(10);50(26);50(50)	*	50(10)
WLSMV	Moderate	*	100(10);50(26);50(50)	*	30(10)
	Severe	*	100(10);50(26);50(50)	*	50(10)

*No no large bias in all conditions.

Under unbalanced moderate non-normality, the ML based estimators still have the same sample size and cluster size recommendation of 50 (10). When severe non-normality is present the ML estimator requires a sample size of 100 for moderate cluster size and sample size of 50 for large cluster size; MLR requires a sample size of 50 with a moderate cluster size of 26. The WLS based estimators have an increased sample size requirement when the clusters are unbalanced; for moderate and severe non-normality it is 100 for small clusters and 50 for medium to large clusters. Table 2.47, below, summarizes the final recommendation. This recommendation gave the sample size recommendation for the lowest cluster size since there seems to be a cluster effect (meaning as cluster size increase, the sample size needed decrease). The balanced versus unbalanced conditions seems to be the same for ML based estimators under moderate non-normality and slightly different under severe non-normality. For the WLS based estimators, the unbalanced moderate normal condition seems to have a bigger sample size requirement and for severe non-normality the sample sizes are roughly the same. Further discussion on this will be discussed within the limitation section.

Table 2.47

Final Sample Size Recommendation

		Final Sample Size Recommendation	
		Moderate	Severe
ML	Balanced	50(10)	100(10);50(26)
	Unbalanced	50(10)	100(26);50(50)
MLR	Balanced	50(10)	100(10); 50(26)
	Unbalanced	50(10)	50(26)
WLSM	Balanced	50(10)	100(10);50(50)
	Unbalanced	100(10);50(26)	100(10);50(26)
WLSMV	Balanced	50(10)	100(10);50(26)
	Unbalanced	100(10);50(26)	100(10);50(26)

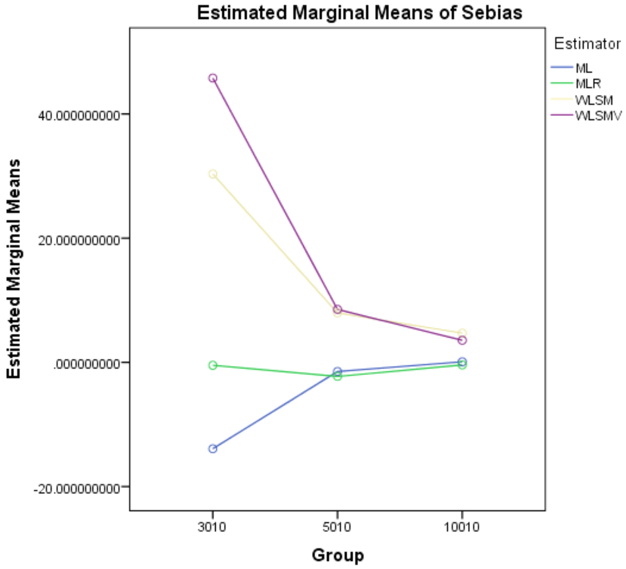
In order to see if there was an interaction marginal means were looked at using SPSS 21.

The first question I asked was *whether the sample size requirement depends on estimation method*. The answer to the question is YES. For the balance moderate normal condition the sample size recommendation was the same but for the unbalanced condition the WLSM estimators required more sample size when the cluster size was low. For severe non-normality, the sample size requirement was the roughly the same for all balanced and unbalanced condition, save for MLR which can needs at least a moderate cluster size. To help me visual what is happening I looked at the marginal means chart to get a clearer picture.

.

Figure 2.8

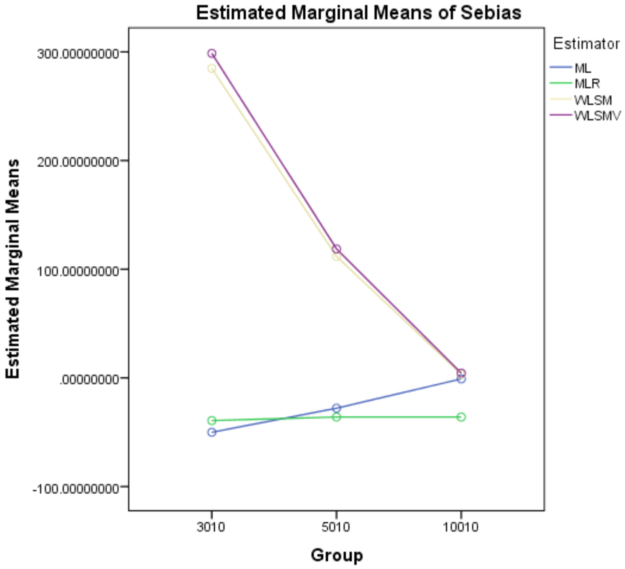
Marginal Mean Graph Estimators by Sample Size (Cluster Ten)



The graph of the marginal or cell means for each group helps determine if there is an interaction. Looking at graph of marginal means for cluster size ten, figure 2.8, there seems to an interaction for the between level SE biases for the ML based vs. WLS based estimators. When the cluster size is low (number of people in the groups is ten), the WLS returns a large positive SE bias. MLR based estimator is fine. ML returns a large negative SE bias. When the sample size is 50 or more the ML based estimators behave very similarly and the WLS based estimators behave very similarly. This means start shows a clear interaction. The mean SE bias for each estimator depends on if the sample size is 30 or 50. There also seems to be a trend the average SE bias seems to get smaller as sample size increase. This is what we want and expect.

Figure 2.9

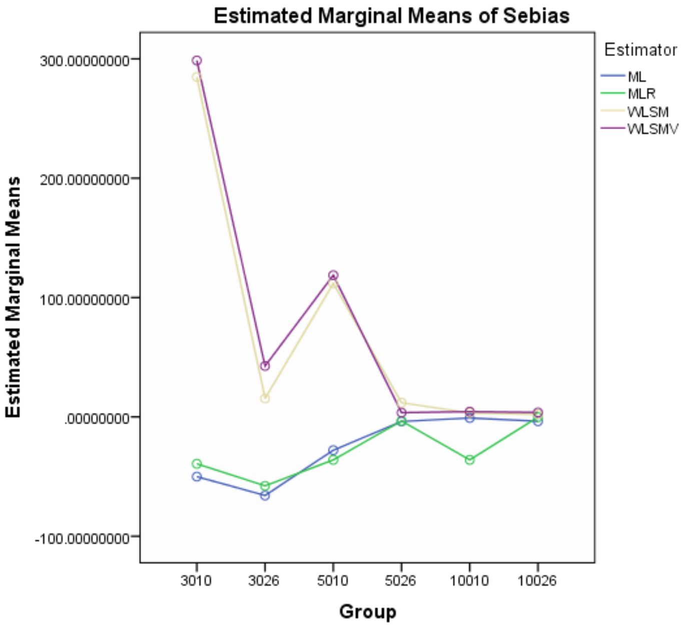
Marginal Mean Graph Estimators by Sample Size (Cluster Ten)-Severe Non-Normality



For severe non-normality, figure 2.9, there is an interaction between sample size and estimators. The WLS based estimators seem to behave similarly and so too does the ML based estimators. The MLR estimator was not recommended when the cluster size was low (ten), it constantly returns large SE biases. The ML based estimator has less of a SE bias when the sample size increases. According to Muthén and Muthén (2002) guidelines a sample size of 100 was needed for ML, WLSM, and WLSMV estimators. There seems to be very little difference in the SE biases for these estimators once the sample sizes reach 100 (under severe non-normality conditions).

Figure 2.10

Marginal Mean Graph Severe Normality Estimator by Multiple Sample and Cluster Size



Putting it all together, figure 2.10, for severely non-normal balanced data, we see that for the ML estimator and WLS based estimators there is stability (no interaction) between sample size 50 cluster size 26 and sample size 100 cluster size 26. Not only are there no interaction the SE bias is very small. For the MLR based estimator it seems a cluster size of at least 26 is needed to establish the same level of stability.

The marginal mean graphs show that there are differences in SE bias based on sample size for the estimators. Although the charts were helpful in seeing what happens as the sample size increase it is not a substitute for the guidelines provided by Muthén and Muthén (2002). My conclusion again is YES the sample size depend on estimator. It also depends on if the estimator is balanced not balanced and severity.

My second research question asked if the sample size requirement greater for normal or non-normal continuous data for the respective estimation method. The primary source for sample size recommendation was based on the research work by Hox et al (2010) that stated that a sample size of 50 was fine for all estimators except MLR which needed a sample size 200 under normal normality conditions.   They found that increasing cluster size or group size had no effect for most estimators except for the pseudobalanced estimation method (MUML). For the MUML estimator, used in their research (not mine), they found increasing group size had a negative effect on the accuracy of the test. In addition, they found whether or not the groups were balanced had no effect and ICC had no effect. Meuleman and Billiet (2009) found an interaction between model complexity and sample size requirement. When the model is simple a sample size of 40 was recommended. My recommendation did somewhat match Hox et al. (2010) sample size recommendation for balanced moderate non-normality. For balanced moderate non-normality, a sample size of 50 was fine when the cluster size was low. A lower number of cluster (or sample size) was sufficient the cluster size increased for the WLS based estimators and ML estimator. For balanced moderate non-normality conditions, MLR required a sample size of 30 minimally. Where Hox (2010) and I disagree: my research found that the within level for all except the MLR estimator had large SE biases, so the between level mattered; being balanced or unbalanced mattered when there was moderate non-normality for the WLS based estimators but under severe non-normality, the recommendations were similar for the balanced and unbalanced groups, balance vs. unbalanced matters; and instead of MLR requiring 200 sample size the sample size was much lower for the MLR based estimator (sample size 30 for moderate non-normal balanced). Since they did not study non-normality, severity was not variable in their research. However, I found severity an important factor.

Note, for unbalanced moderate non-normality, the cluster size requirement for WLS based estimators is important. In order to have a sample size of 50, a moderate cluster size of 26 is needed for unbalanced moderate non-normal WLS based estimators. When severe non-normality is present, for balanced small clusters sizes, a sample size of 100 is required.

However, most estimators required a moderate cluster size of 26 with sample size of 50 (WLSM required a large cluster size of 50). When unbalanced severe non-normality was present the WLS based estimators required a sample size of 100 for small cluster size and sample size of 50 for moderate (CS =26) cluster size. For ML based estimators unbalanced severe non-normality, when the cluster size is small, a sample size of well over 100 is required. For MLR sample size of 50 will require a moderate cluster size and for ML a sample size of 50 would require a large cluster size. Again, sample size recommendation seems to depend on severity, estimator, and whether balanced or unbalanced.

My last question was to determine if the presence or absence of unbalanced clusters affected the sample size requirement. The previous paragraph says yes the presence of balanced or unbalanced clusters did affect what was the sample size requirement. Moderate non-normal WLS based estimators required a larger sample size for when small cluster size was present (the ML based estimators were not affected). Under severe non-normality, the ML based estimators unbalanced condition generally required either greater sample size or cluster size. The WLSM cluster size recommendation of 50 for its balanced severe non-normality condition might be because of small number of replications done to stabilize the standard errors (see limitations). I suspect the sample size requirement under severe-non normality is unaffected by whether or whether not my conditions are balanced.

Conclusion

I will attempt to explain why we are seeing what we are seeing in my final thoughts. According to Brown (2006), ML is sensitive to excessive kurtosis and tend to underestimate SE of the parameter (increase type I error). He recommends MLR and WLS based estimators when non-normality exist but WLS based estimators require a large sample size and MLR based is good but only when floor and ceiling effect is not present. Here we saw that when the overall data are slightly non-normal, that that slight non-normality still extended to the between level. MLR performed very well but ML and WLS based estimators required a higher sample size with WLSM based estimators generally requiring greater sample size requirement. The ML estimator did tend to underestimate the SE bias while the WLS estimator tended to overestimate the bias when the sample size (or the number of clusters) was low. Over all the conditions of non-normality, MLR tended to perform best both at the between and within levels and most times had a lower sample size requirement. The sample size requirement for MLR estimator, although low, was affected by non-normality severity and whether or not the clusters were balanced for moderate non-normality conditions. Future research would need to explore why some researchers have found no problems at the within level and other have found problems.

Limitations

A drawback of Monte Carlo results is that they are conditional on the design, and generalizations are therefore only justified when there is a clear trend (Hoogland and Boosma, 1998, p. 330). My design was a simple design and I could not cover the gamut of all sample sizes, cluster sizes, and structures in this multilevel model. This study attempted to give some basic insight on sample requirement when non-normality is present and variables that could be interwoven. Another limitation was the number of replications and its' affect on the standard errors. Although Hox et al. (2010) used 1000 reps to do their simulation, 1000 reps might not have been too low. According to Bandalos (2006) when non-normality is present the standard errors tend to bounce. Although no guideline was given, at least 10,000 (or as high as 100,000 or more) might be needed to stabilize the standard errors. In this study, the standard errors on average tended to be corrected (so I was confident in the decision drawn) but the individual indicators standard errors varied more than necessary, and the subsequent rerunning of the conditions sometimes yield different numbers but with the same conclusions. Because of the increase computer time (and the FORT 16 out of memory errors), it was difficult to run numerous conditions with a large number of replications, 100k. Each attempt to try resulted in Unix server errors and my personal computer crashing. However, I did run one condition for 100,000 reps, namely, severely non-normal ML balanced sample (cluster) size 30 (10) and obtained average relative standard error bias of -46.13 vs. at 1000 reps the average relative standard error bias for this condition was -50.07. Not only were the biases smaller, but the MSE (mean square error) descended from 2.85 to 2.612. The conclusion remained the same.

CHAPTER SIX

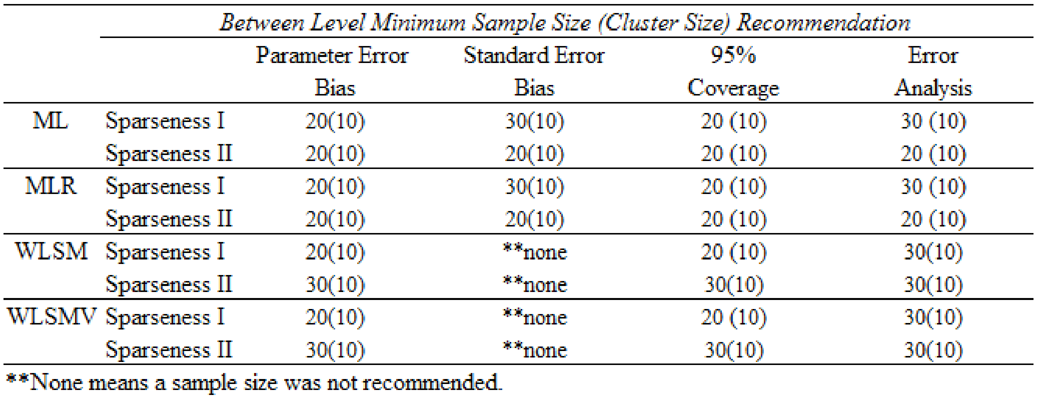
STUDY THREE CONCLUSION

Study three we studied the sample size requirement when the data are categorical and when sparse conditions are present. Condition I had more sparseness in the cells (Sparseness I) while condition II had less sparse cells (Sparseness II). Table 2.48, below, shows one of the overall sample size recommendation for study three. For the ML based recommendation, sparseness I condition, the sample size recommendation based on the Muthén and Muthén (2010) requirement was 30(10) --if we look at the standard error bias and/or error analysis bias. It seems that for sparseness II the sample size (cluster size) recommendation would be 20 (10). However, it took such a long time to run --much more than 30 (10) --that for practicality a sample size of 20 might still be too low even for sparseness I. For the WLS based estimators because the standard errors tended to increase as the sample size increased it is not recommended that one use WLS based estimators at the between level when the data are considered categorical.

Table 2.48, also below, shows the sample size recommendation for the unbalanced condition. Although I used a Unix server to run multiple conditions each account is only allowed a few number of processors at one time, so because the first set of conditions (balanced) took so long to converge (almost a month) the second batch (unbalanced) started late and did not have a chance to converge. The results are the balanced and unbalanced conditions seemed to be very similar, so I doubt there would have been a difference in the recommendation. For the unbalanced condition the sample size (cluster size) recommendation is 30 (10) for the ML based estimators and there are no sample size recommendations for the WLS based estimators.

Table 2.48

Study 3 Balanced Between Level Minimum Sample Size Recommendations

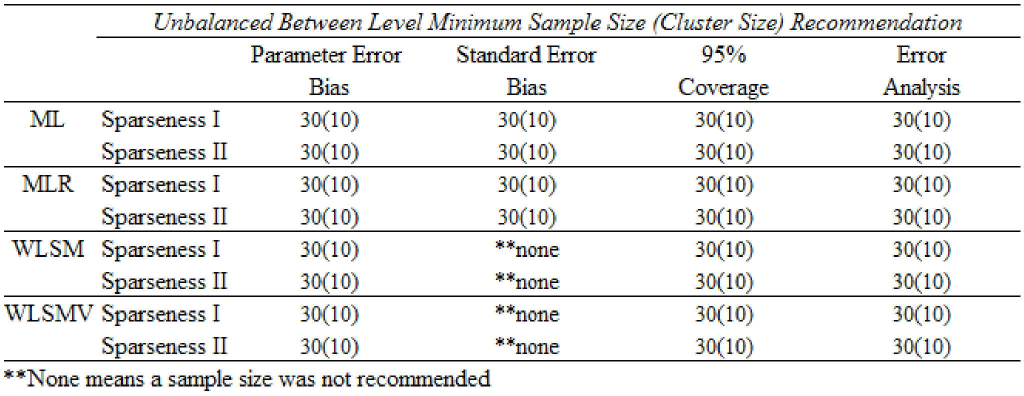


		Between Level Minimum Sample Size (Cluster Size) Recommendation			
		Parameter Error Bias	Standard Error Bias	95% Coverage	Error Analysis
ML	Sparseness I	20(10)	30(10)	20 (10)	30 (10)
	Sparseness II	20(10)	20(10)	20 (10)	20 (10)
MLR	Sparseness I	20(10)	30(10)	20 (10)	30 (10)
	Sparseness II	20(10)	20(10)	20 (10)	20 (10)
WLSM	Sparseness I	20(10)	**none	20 (10)	30(10)
	Sparseness II	30(10)	**none	30(10)	30(10)
WLSMV	Sparseness I	20(10)	**none	20 (10)	30(10)
	Sparseness II	30(10)	**none	30(10)	30(10)

**None means a sample size was not recommended.

Table 2.49

Study 3 Unbalanced Between Level Minimum Sample Size Recommendations



		Unbalanced Between Level Minimum Sample Size (Cluster Size) Recommendation			
		Parameter Error Bias	Standard Error Bias	95% Coverage	Error Analysis
ML	Sparseness I	30(10)	30(10)	30(10)	30(10)
	Sparseness II	30(10)	30(10)	30(10)	30(10)
MLR	Sparseness I	30(10)	30(10)	30(10)	30(10)
	Sparseness II	30(10)	30(10)	30(10)	30(10)
WLSM	Sparseness I	30(10)	**none	30(10)	30(10)
	Sparseness II	30(10)	**none	30(10)	30(10)
WLSMV	Sparseness I	30(10)	**none	30(10)	30(10)
	Sparseness II	30(10)	**none	30(10)	30(10)

**None means a sample size was not recommended

Like study two my analysis was based on Muthén and Muthén (2002) list of requirements for determining sample size. The four research questions were as follows:

(1) Does sample size requirement for categorical independent variable data depend on estimation method?

(2) Is the sample size requirement the same or different compared to the normal multilevel data for the respective estimation method?

(3) Does the presence or absence of unbalanced clusters affect the sample size requirement for categorical data?

(4) Does the presence of sparse tables affect the sample size requirement?

Does the sample size requirement depend on estimation method? Or *Does sample size requirement for categorical independent variable data depend on estimation method?* This question seeks to ascertain if the sample size requirement is the same for all estimation methods. I looked at Muthén and Muthén (2002) guideline to help answer this question; also, like in study two, I looked at the marginal means chart. Doing standard statistical analysis was a moot point since sample was not one of variables in the data set (specifically, not the dependent variable), I had to answer this question in terms of the between level standard error bias as it measures up to the Muthén and Muthén (2002) guidelines. Again, sample size twenty was only a test case and not included in the analysis portion since this sample size did not fully run and there were a good number of non-convergent conditions. The sample size starts at thirty. Again, using SPSS 21, I was able to look at the marginal means.

Figure 2.11

Means Plots Sample Size 30 (Cluster Size 10)

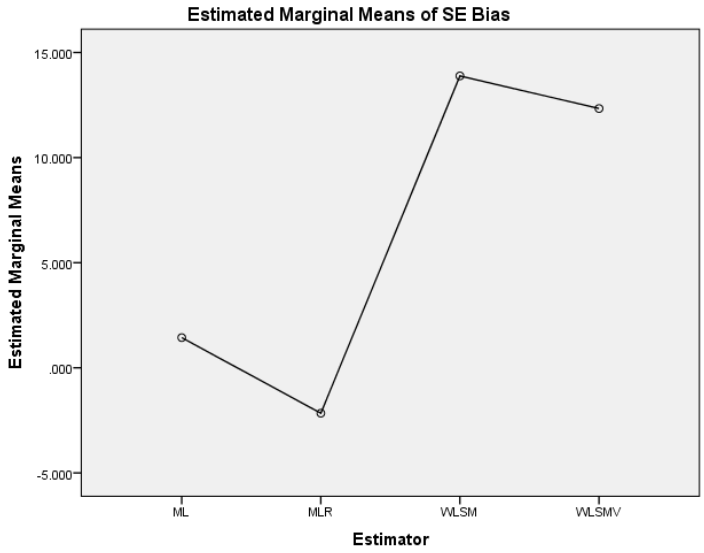
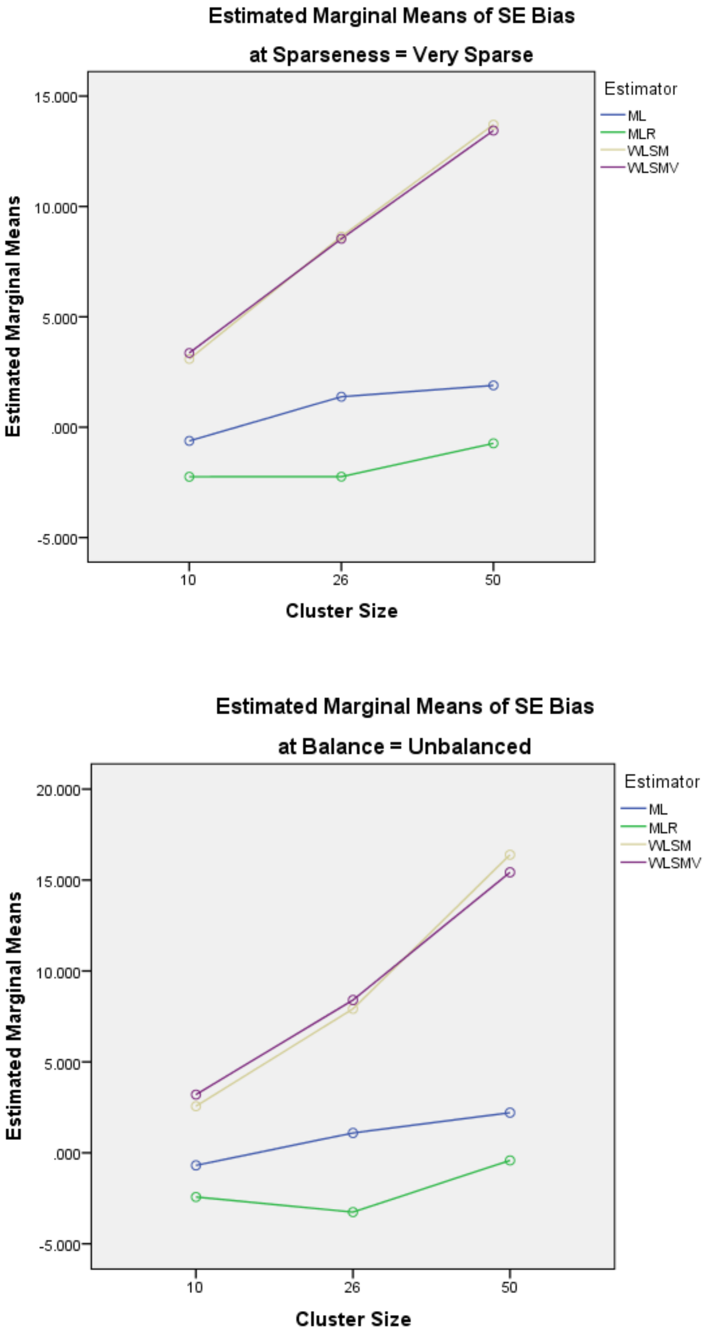


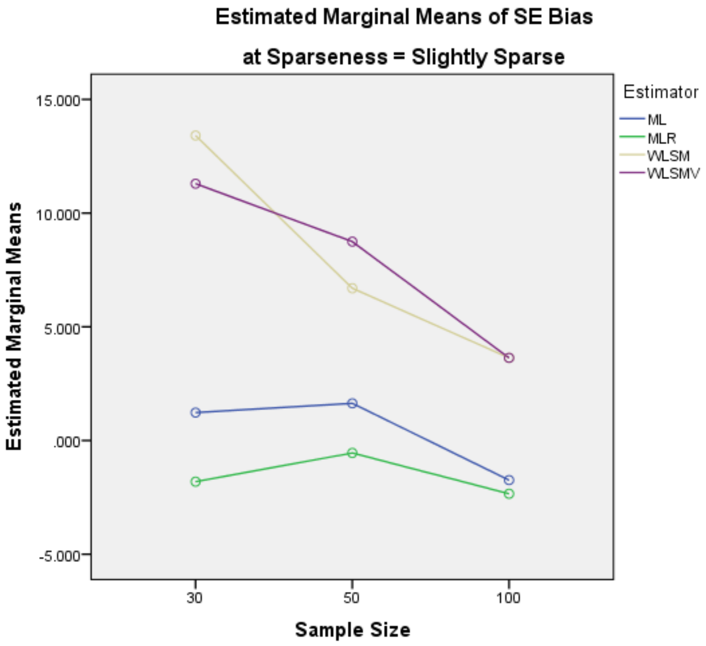
Figure 2.11 is a general graph of the marginal means across all estimators when the sample size is thirty and cluster size is ten. We looked at the 30(10) sample size (cluster size) condition because the sample size recommendation was 30 for the ML based estimators and not recommended for the WLS based estimators. As you can see the SE bias for the ML estimators were not that different from one another, but the WLS estimators the between level SE bias were larger (more positive) than the ML based estimators and exceeded our 10% guideline. So the answer to the question is an emphatic YES it does. The ML based estimators require a sample size of 30 and there was no sample size recommendation for the WLS based estimators.

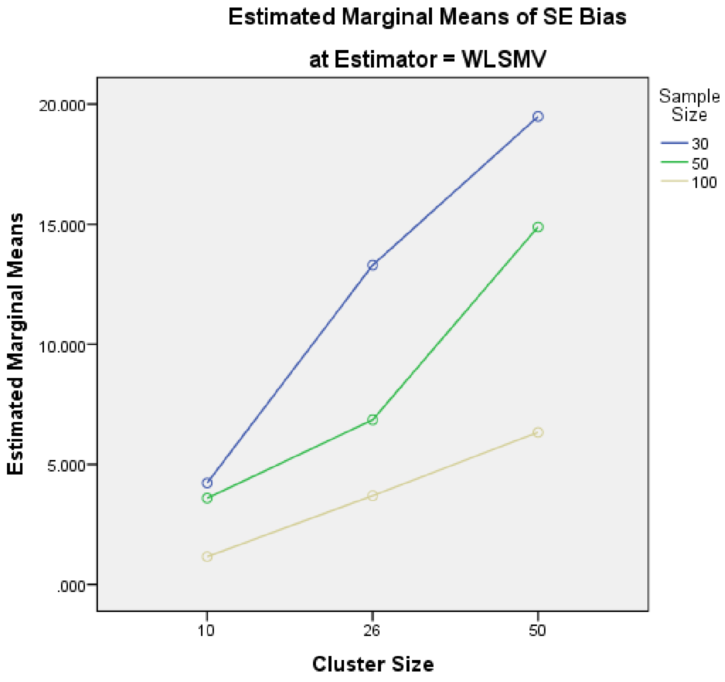
Further exploration was needed due to the behavior of WLS based estimators. If you recall from the experiment as cluster size increase the standard errors for the WLS based estimators increased; in fact, the sample size of thirty was sufficient as long as the cluster size was low but not so when the cluster size increased for the WLS based estimators. Here, at sample size thirty and cluster size ten, the ML based estimators are not significantly different from each other and the SE bias is under the Muthén and Muthén (2002) criterion but with WLS based estimators as the clusters size increase at sample size thirty so did the bias. There may be an interaction between cluster size and sample size for the WLS based estimators.

Figure 2.12

Further Exploration



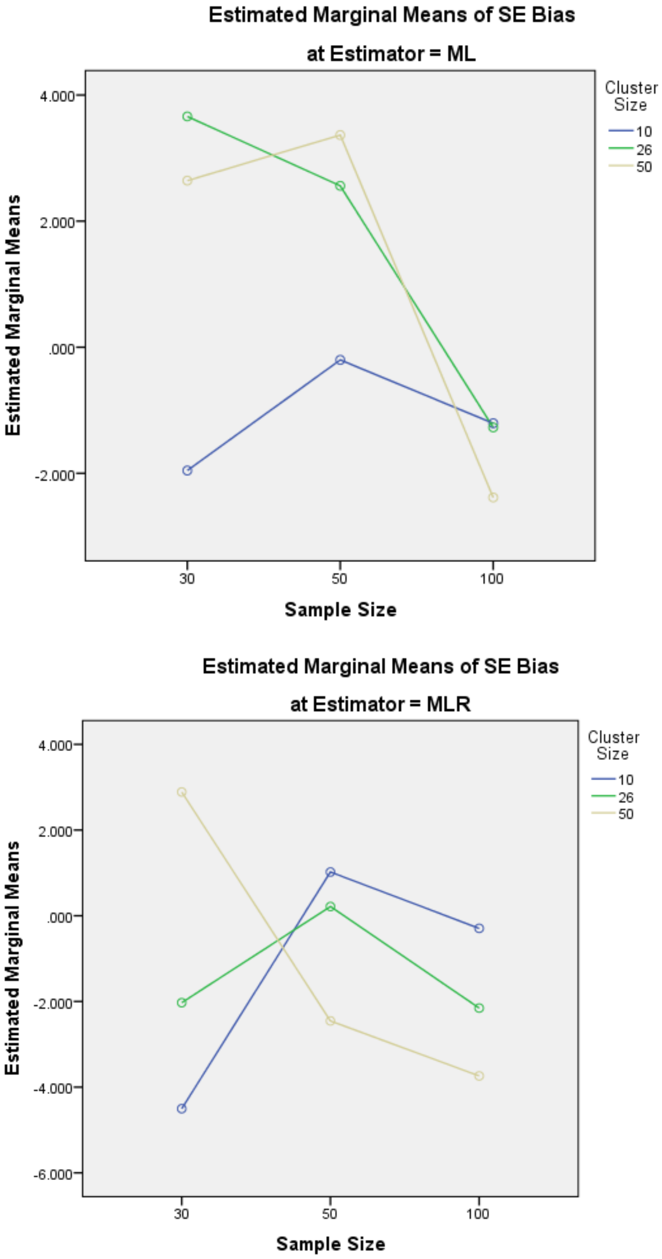


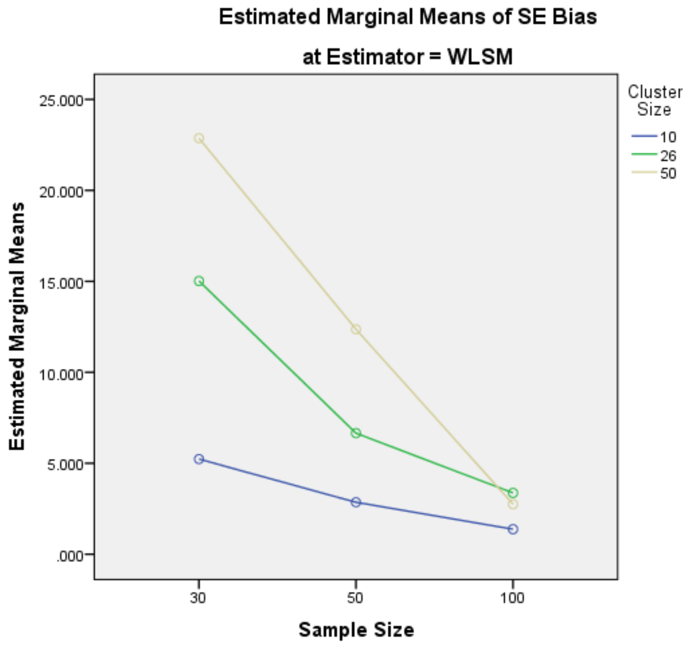


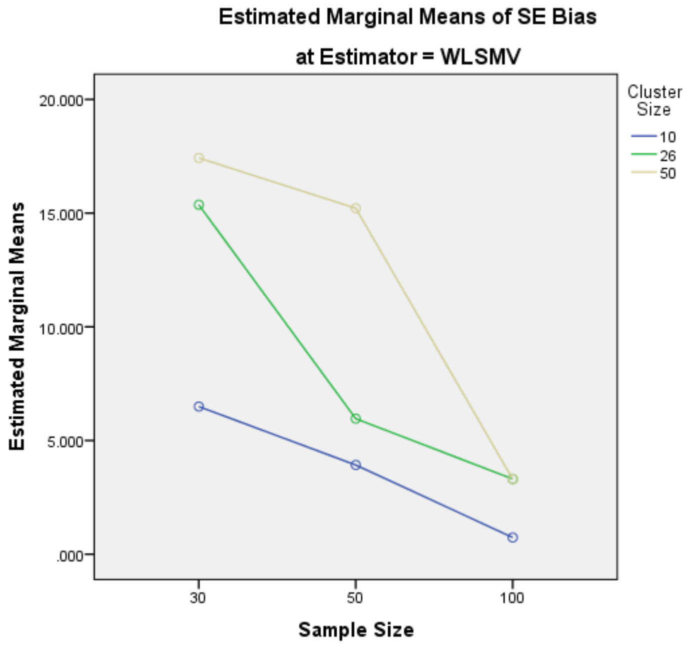
The various figures under figure 28 show that regardless of balance or sparseness the WLS base estimators' SE bias seem to increase as cluster size increase. As sample size increase the SE bias decrease for all estimators regardless of cluster size. The last chart shows without loss of generality, without loss of generality (WLOG), that there is an interaction between cluster size and sample size.

Figure 2.13

Three way Marginal Means Plot, Sample Size*Cluster*Estimator for Balanced Sparse I Data



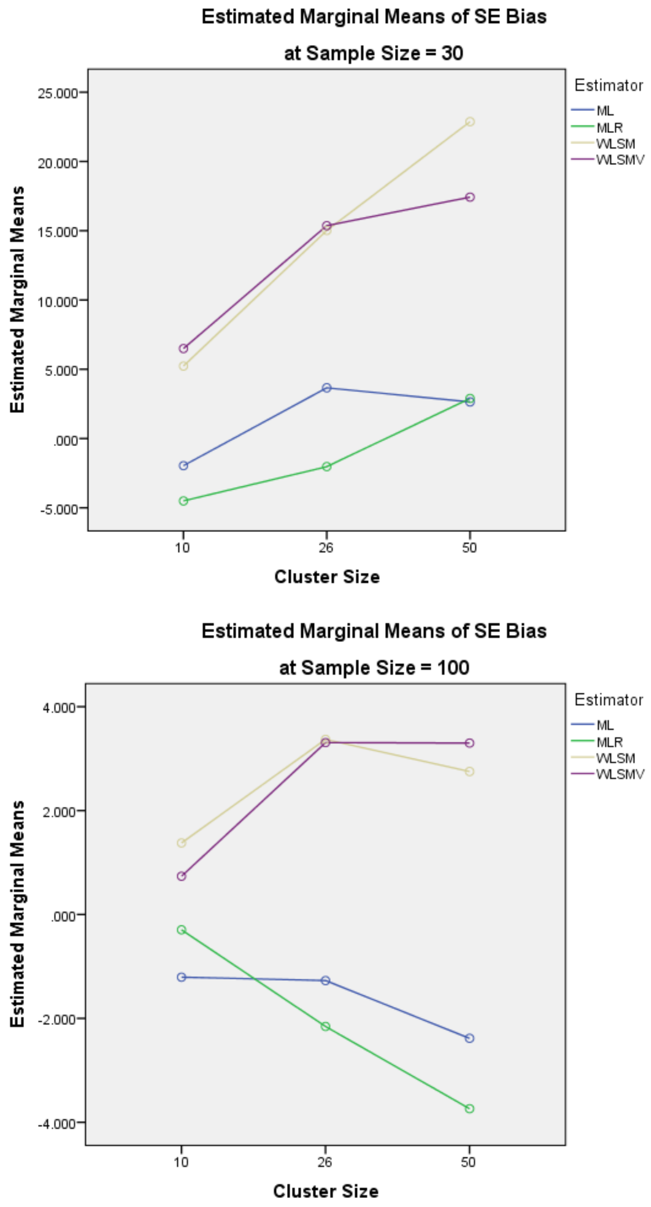




There seem to be an interaction between sample size and cluster size for each of the estimators, and they all seem to be different. For the WLS based estimators, an appropriate SE bias occurs the sample size is 100 for all three cluster sizes. A sample size of 100 might be needed for WLS estimators; however, this is only for balanced very sparse data. Investigation will have to be done to see if this pattern holds for all four conditions. A sample size of 30 is still sufficient for ML based estimators.

Figure 2.14

Three way Marginal Means Plot, Sample Size*Cluster*Estimator for Balanced Sparse I Data

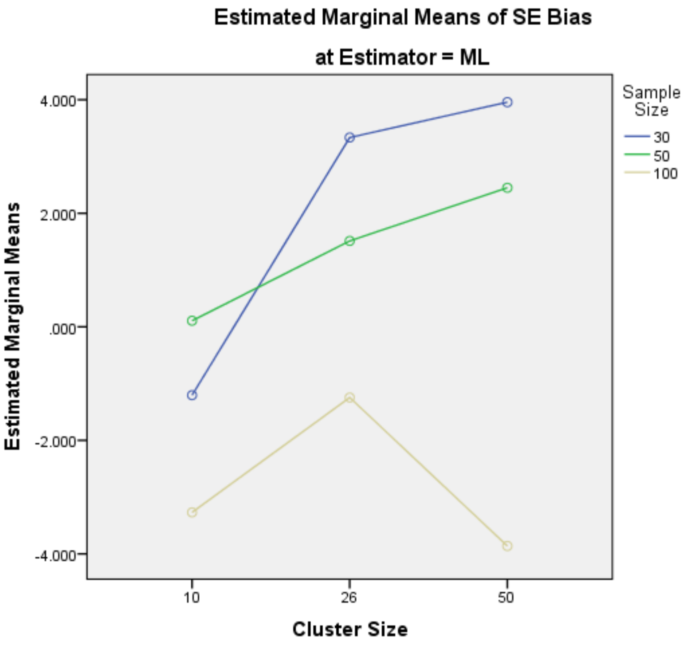


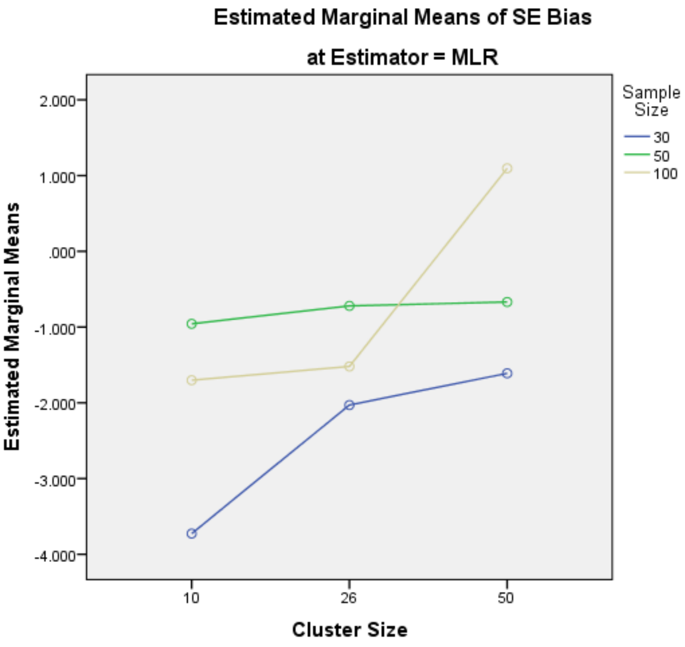
When the sample size and cluster size 100(10) there was the difference between estimators' SE biases is very slight but as the cluster size became bigger there was a bigger difference between estimators' SE biases. Note, in for all four estimators they were within an acceptable range to recommend the sample size at all cluster levels. At 100(50), the WLS based estimators seemed to level off, as far as the bias was concerned. The upward trend seemed to dissipate. It appears a 100 sample size would be recommended for the WLS based estimators, but this is only for the balanced very sparse case. Despite the trend seen, the sample size recommendation, again, does depend on estimation method used, and as we see it also depends on the cluster size.

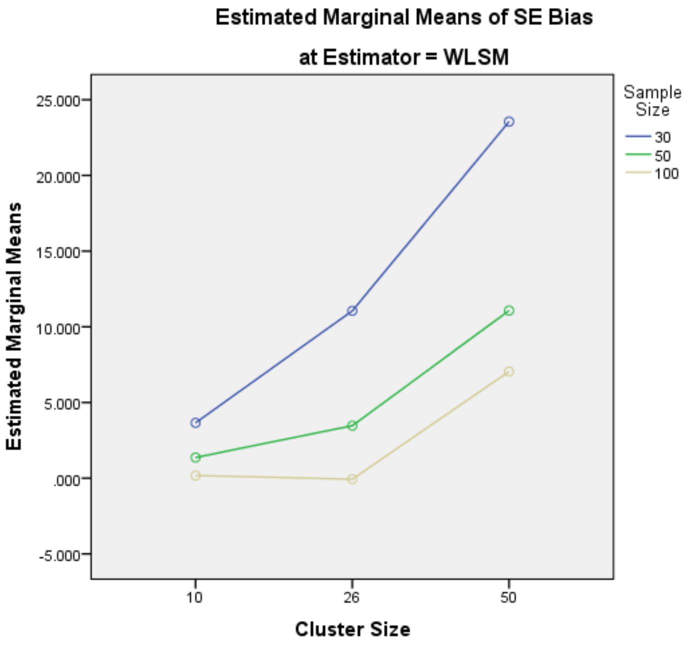
The second condition was sparseness level two (*not very sparse*) and balanced condition. Figure 2.15 demonstrates a three way interaction between estimator, cluster size, and sample size. As with the previous condition, as sample size increased to 100 for the WLS based estimators, the bias decreased in general but the standard error bias increased as cluster size increased. The average standard error bias was definitely under the 10% Muthén and Muthén threshold at sample size 100(10) and still under at 100 (50). It should be noted that the standard error bias seem to still increase even at cluster size 50 so it the cluster size increased beyond 50 then sample size 100 might not be sufficient.

Figure 2.15

Three way Marginal Means Plot, Sample Size*Cluster*Estimator for Balanced Sparse II Data


Estimated Marginal Means of SE Bias
at Estimator = ML
Estimated Marginal Means
4.000
2.000
.000
-2.000
-4.000
10
26
50
Cluster Size
Sample Size
30
50
100


Estimated Marginal Means of SE Bias
at Estimator = MLR
Estimated Marginal Means
2.000
1.000
.000
-1.000
-2.000
-3.000
-4.000
10
26
50
Cluster Size
Sample Size
30
50
100



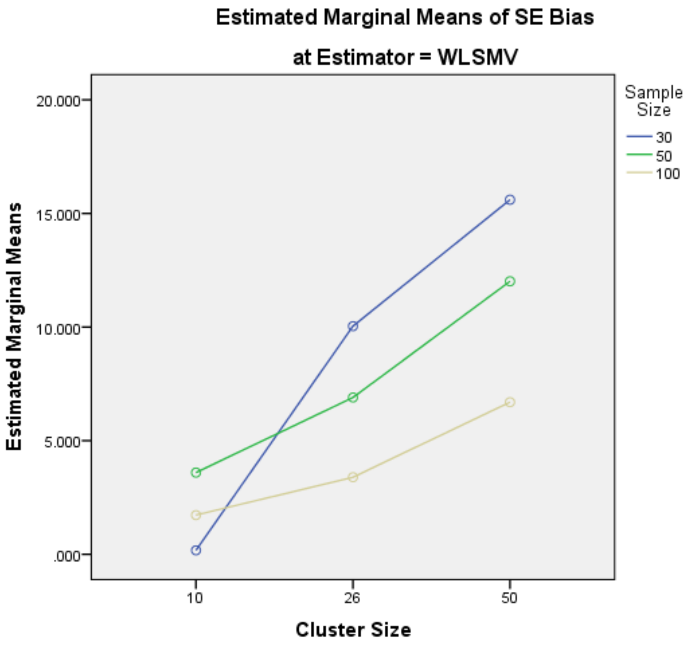
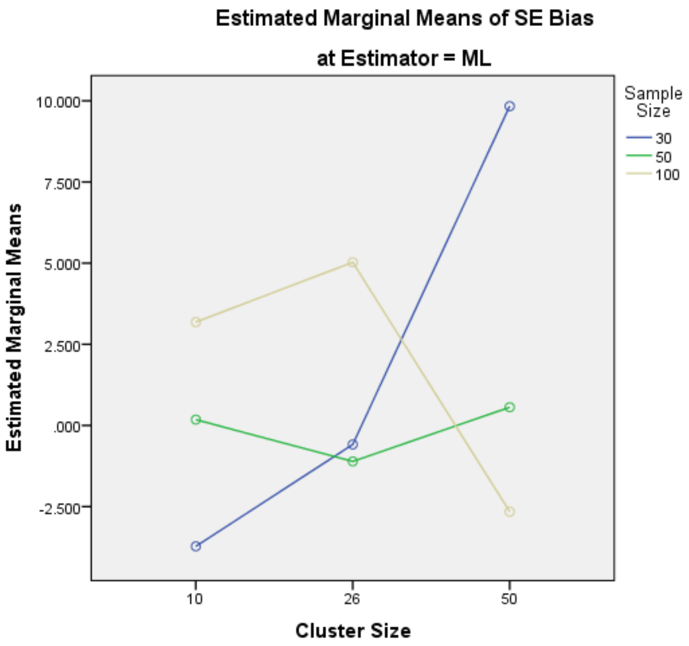


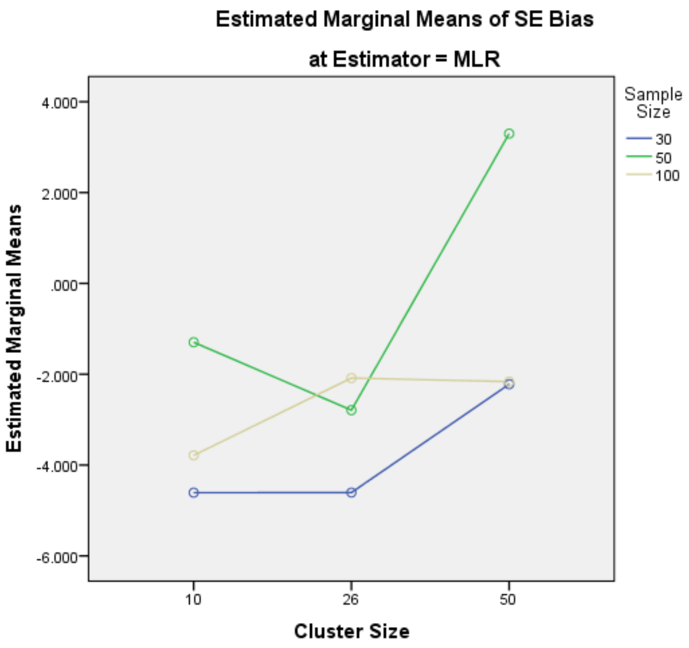
Figure 2.15 shows the marginal means plot for very sparse and unbalanced data. Looking at Figure 2.15 the usual sample size recommendation of 100 barely works for the WLSM estimator because the SE bias is over the 10% bias guideline; WLSMV is within the guideline.

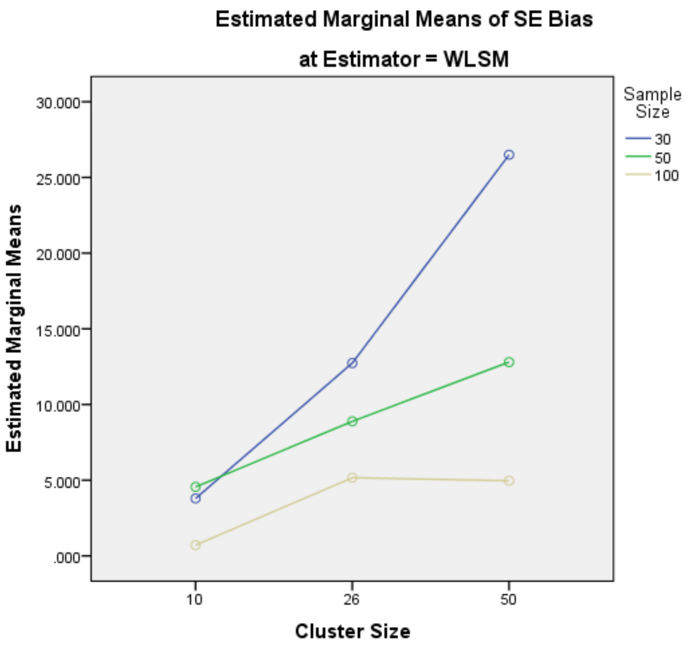
Also, the ML based estimator had a large (although within guidelines) increase in SE bias as sample size increased. The ML based estimators pattern of behavior was less obvious than WLS based estimators. However, it is clear that all estimators show an interaction between cluster size and sample size. What should be noted is that in this study I only went up to a cluster size of 50, but if the cluster size was 100 would a sample size of 100 be recommended for the WLS estimators? Would that larger cluster size also affect the ML based estimators as well? This study is not all encompassing but meant to start a much needed conversation. There may need to be a ratio of sample size to cluster size needed because of the interaction that seem to exist using the marginal means charts.

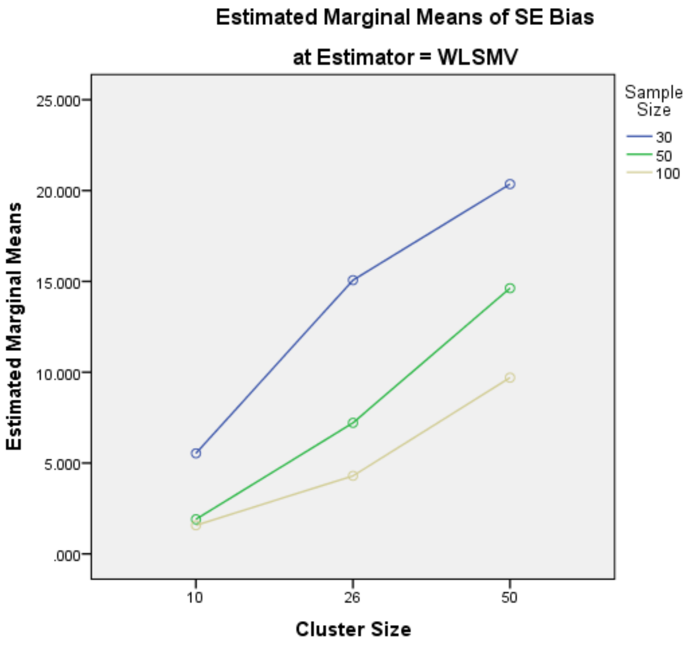
Figure 2.16

Three way Marginal Means Plot, Sample Size*Cluster*Estimator, Unbalanced Sparse I Data




Estimated Marginal Means of SE Bias
at Estimator = MLR
Estimated Marginal Means
4.000
2.000
.000
-2.000
-4.000
-6.000
10
26
50
Cluster Size
Sample Size
30
50
100


Estimated Marginal Means of SE Bias
at Estimator = WLSM
Estimated Marginal Means
30.000
25.000
20.000
15.000
10.000
5.000
.000
10
26
50
Cluster Size
Sample Size
30
50
100

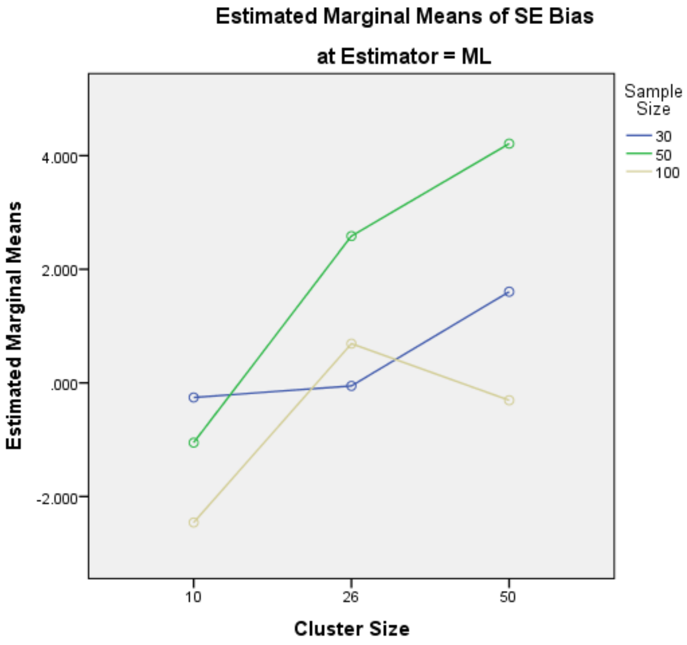


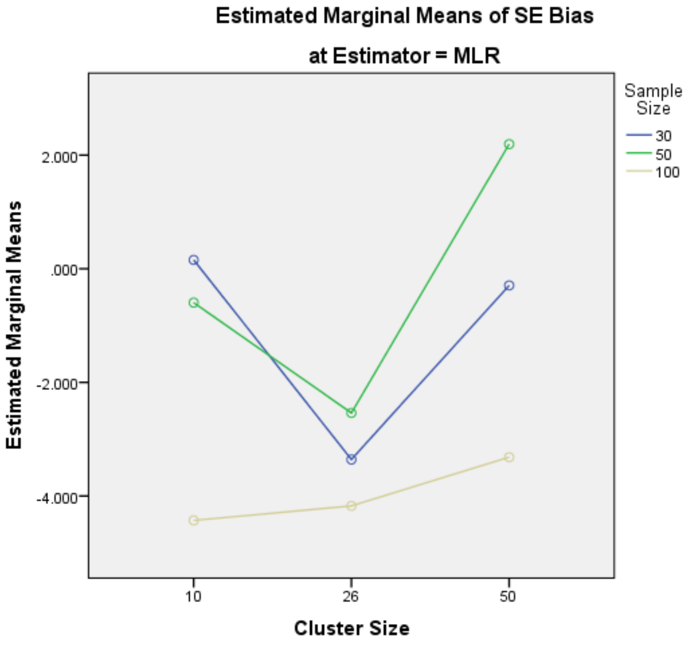
Looking at the obvious patterns within the marginal means charts, we see despite the clearly elevated SE bias for the ML and WLSMV estimators at the recommended sample size (cluster size), the recommended sample size of 30 (10) is okay for the ML based estimators and 100(10) for the WLS based estimators (well, if the cluster size is small).

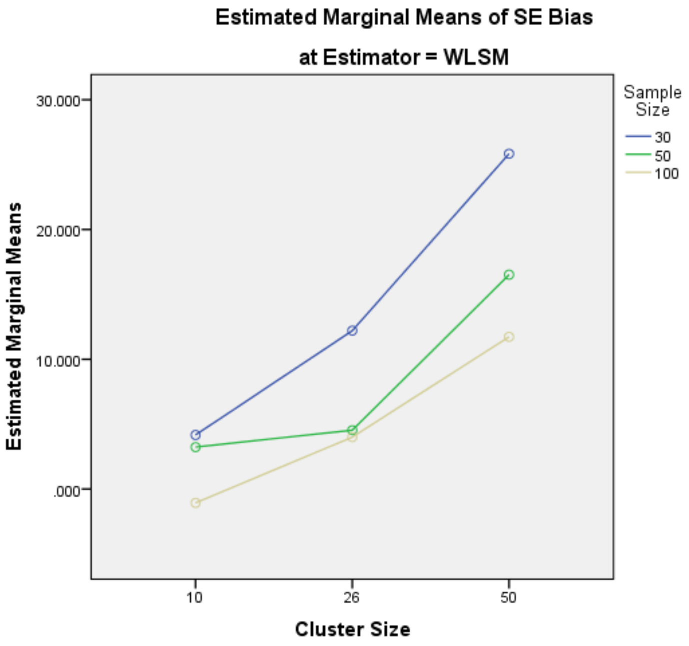
The last condition was the unbalanced with only slightly sparse data, sparse two. One can look Figure 2.17 and see that each of the interactions look differently across estimators and within each estimator there are several non-parallel lines indicating a two way interaction exist as well at each level of the estimator variable. Across all estimators there when the sample size was 100 the SE bias was within the Muthén and Muthén (2002) guidelines but the WLSM estimator SE bias seems to borderline and increasing. Again, if my study included larger cluster sizes then the 100 sample size recommendation might not actually hold.

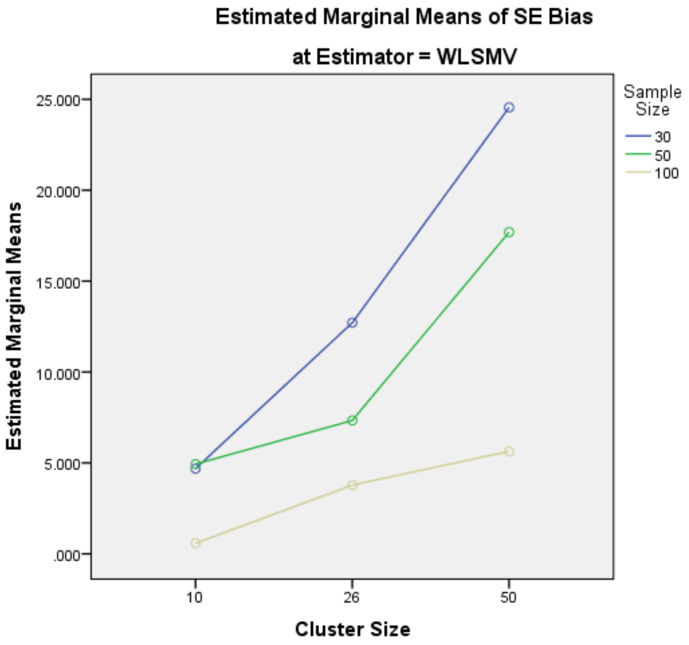
Figure 2.17

Three way Marginal Means Plot, Sample Size*Cluster*Estimator, Unbalanced Sparse II Data









For lightly sparse unbalanced data, WLSM was over the SE bias guideline and for very sparse unbalanced data WLSMV was large but slightly under the SE bias guideline. For balanced data there was no difference between the two estimators when the sample and cluster size were high. For this condition a sample size cluster size of 30 (10) seem to be sufficient for ML based estimators and 100 for WLS based estimators (there were still reservations about this recommendation).

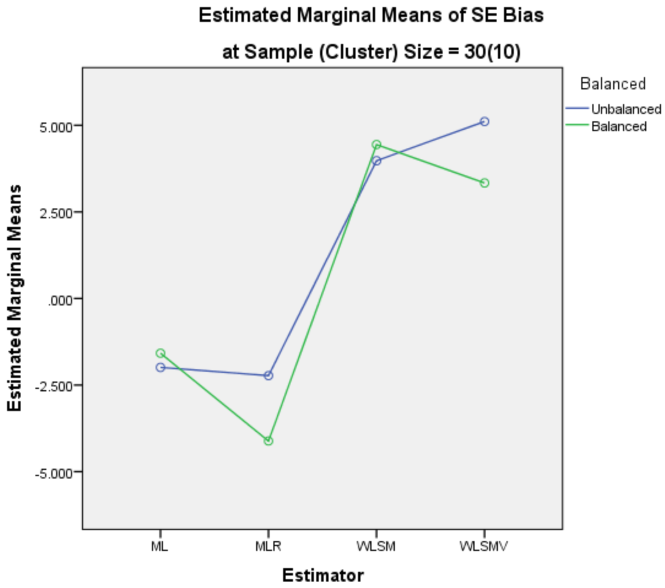
So, *my answer to the question is the sample size recommendation different depending on estimation method*, as I said before, yes! In general the ML based estimation method requires a sample size of 30 and the WLS estimation method for this experiment needed a sample size of at least 100. However, because there seems to be a clear interaction between sample size and cluster size especially when for WLS estimator, it appears that there may be a need for a golden ratio that could state the sample size relative to cluster size since there is an increasing pattern for the most part.

The second *question asked if the sample size requirement the same for normal continuous multilevel data.* The answer is, NO. When the data are normal Hox et. al (2010) found a sample size of 50 was sufficient for WLS based estimators and ML estimator while MLR needed at least 200 groups/sample size to have adequate coverage. The sample size requirement for the WLS based estimators depends on the cluster size. The ML based estimators were sufficient at sample size 30.   Also, the sample size recommendation was lower than the recommendation non-normal continuous data.

The third question asked if *the presence of balanced or unbalanced data affected the sample size.* For WLS based data the presence of unbalanced data did affect the sample size. Looking at Figures 2.18-2.20, I saw a clear three way interaction and a two way interaction. The graphs were slightly different for each sample size (cluster size). What was clear in the figures was that for the WLS based estimators the unbalanced SE bias was almost always higher. What was even clearer was when the cluster size was 50 (large), there was a main effect between unbalanced data and balanced data especially for the WLS based estimators. When the cluster size was 50 (large) for the WLS based estimators the SE bias was always larger for the unbalanced data, Figure 2.21. Although it was difficult to recommend a sample size for WLS based estimators, having the extra complication of unbalanced data seemed to cause even more SE bias.

Figure 2.18

Interaction between Balanced and Estimation for Sample Size 30, Study3 data



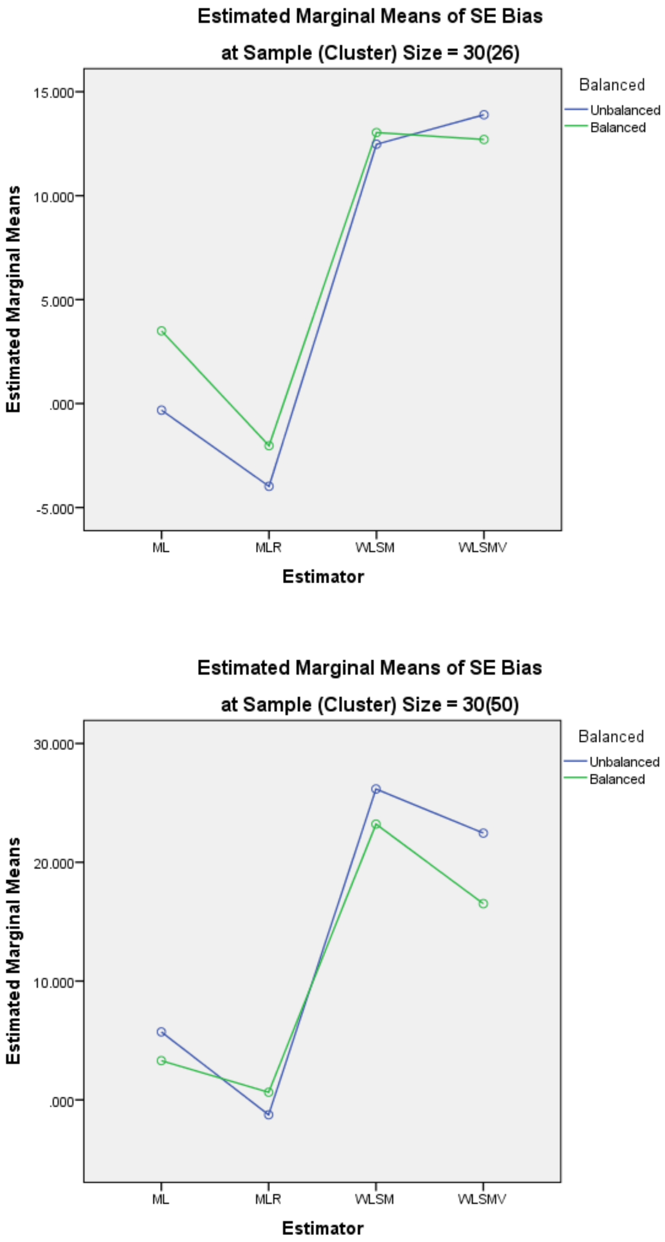
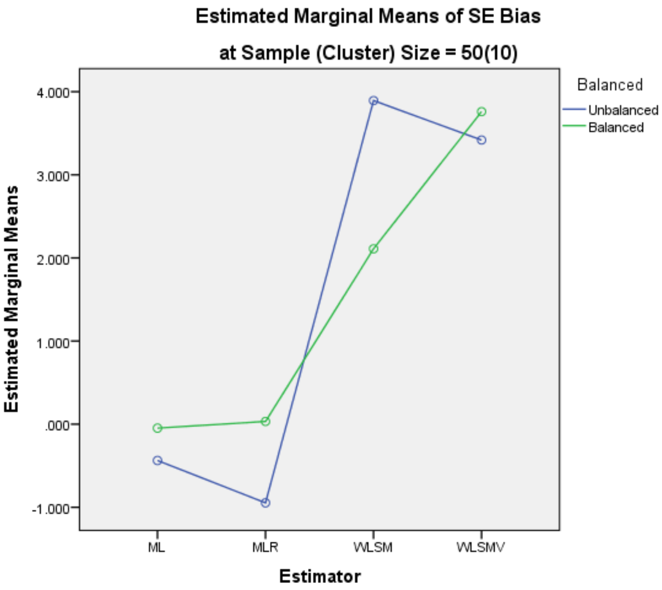
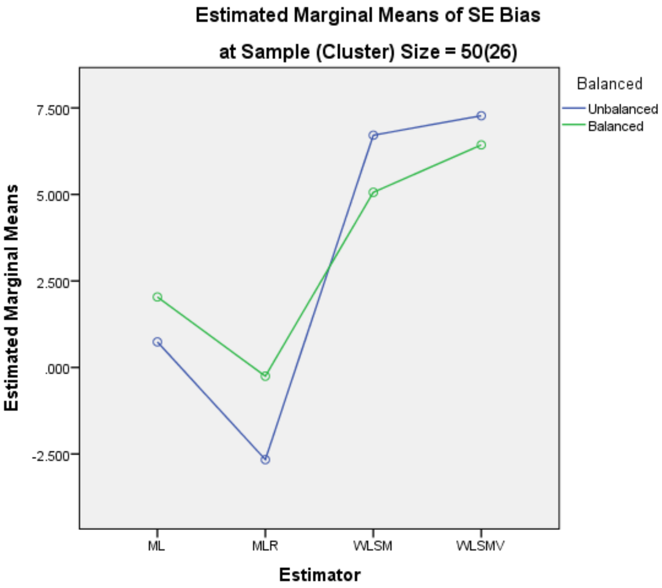


Figure 2.19

Interaction between Balanced and Estimation for Sample Size 50, Study3 data


Estimated Marginal Means of SE Bias
at Sample (Cluster) Size = 50(10)
Estimated Marginal Means
4.000
3.000
2.000
1.000
.000
-1.000
ML
MLR
WLSM
WLSMV
Estimator
Balanced
Unbalanced
Balanced


Estimated Marginal Means of SE Bias
at Sample (Cluster) Size = 50(26)
Estimated Marginal Means
7.500
5.000
2.500
.000
-2.500
ML
MLR
WLSM
WLSMV
Estimator
Balanced
Unbalanced
Balanced

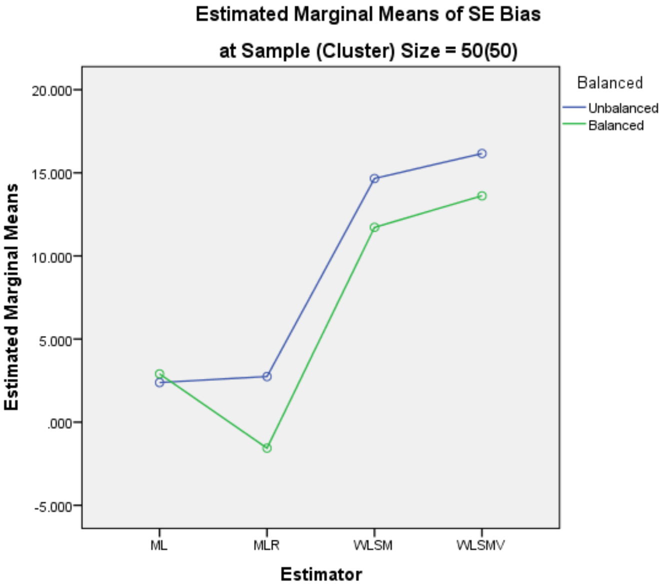
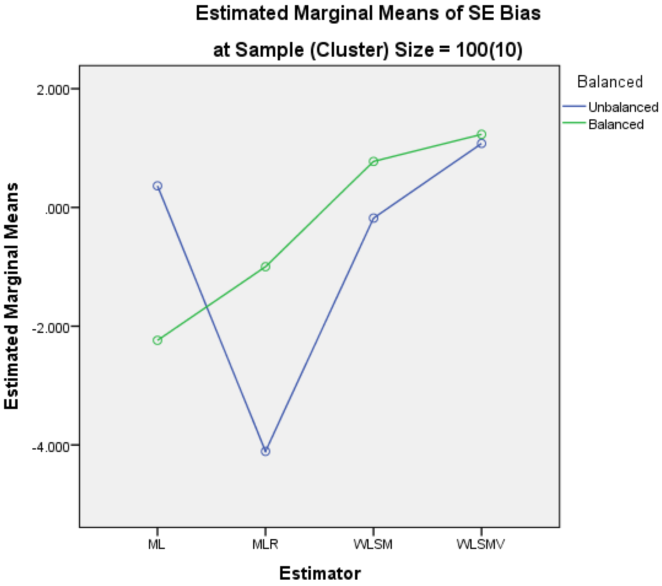
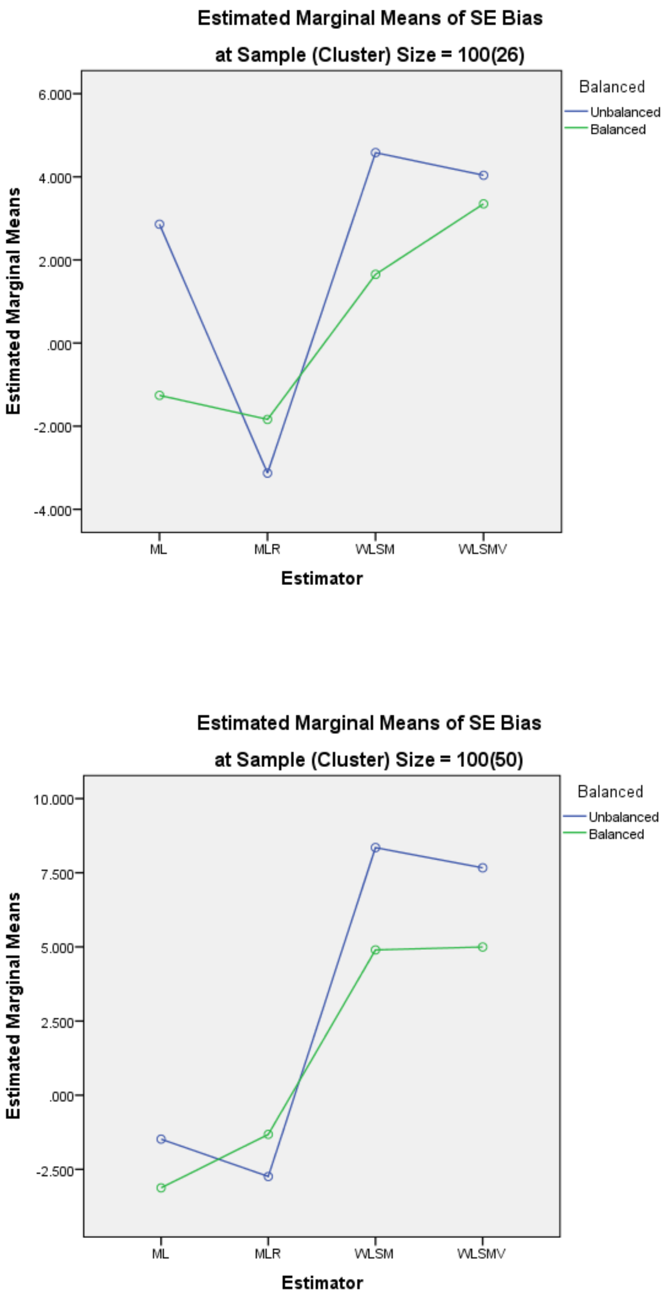


Figure 2.20

Interaction between Balanced and Estimation for Sample Size 100, Study3 data



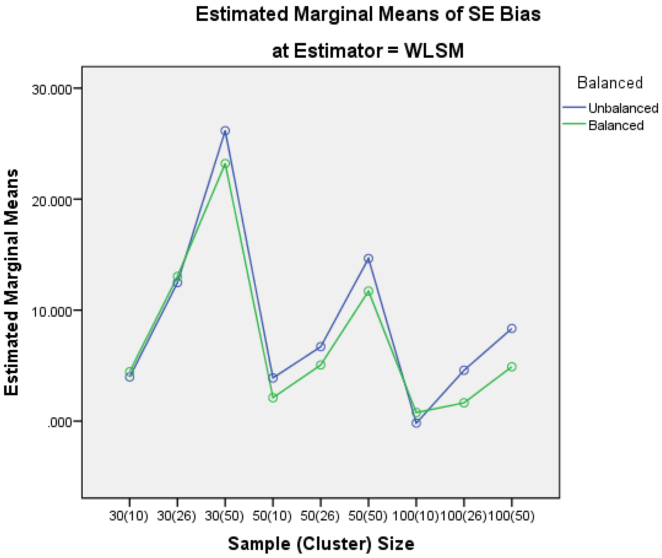


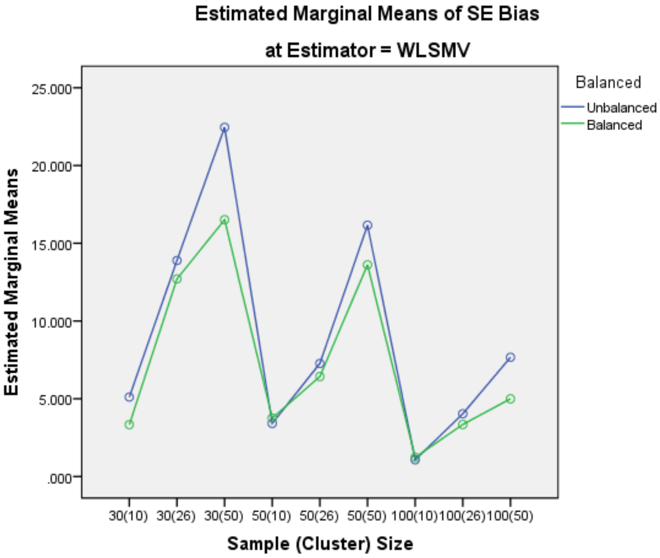
Estimated Marginal Means of SE Bias at Sample (Cluster) Size = 100(26)

Estimated Marginal Means of SE Bias at Sample (Cluster) Size = 100(50)

Figure 2.21

Interaction between Balance and Sample size for the WLS estimators, Study3 data



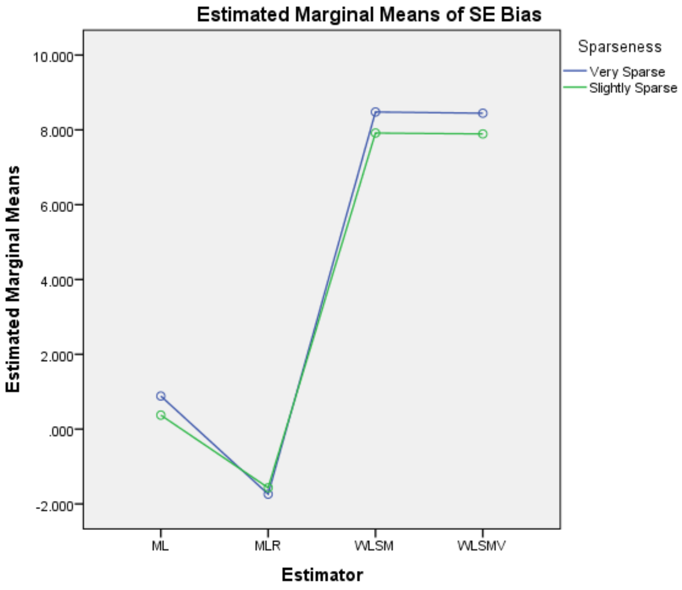


To answer the question, since I am not giving a recommendation for the WLS based sample size, and there were no really clear pattern observed for the ML based estimators, then *there was no a difference between sample size recommendations when the group was balanced versus unbalanced*.

Lastly, *did the presence of sparse tables affect the sample size requirement?* No or nil, from what I saw. Figure 2.22 show that there is a main effect for WLS based data and no main effect and very little of an interaction present for ML based data.

Figure 2.22

Interaction of estimation and sparseness, study3



In conclusion, for ML based estimators a sample size of 30 is fine except for WLS based estimators even though in most cases a sample size of 100 was okay. There did seem to be a trend due the interaction seen between sample size and cluster size for the WLS based estimators (which may exist for the ML based estimators but to a smaller extent). The larger the cluster size, the larger the sample size must be for the WLS based estimators. For most conditions in my study except under unbalanced case I was able recommend a sample size of 100. There may be a golden ratio but that is outside the scope of this study. Based on theory I expected the WLS based estimator to perform extremely well on categorical data but according to Brown (2006) the sample requirement is large and is a function of the model size. For a moderate size CFA model of 10 to 15 indicators, a sample size of 150 to 200 was needed for the WLSMV estimator (Brown, 2006, p.389). The ML estimator generally is not recommended for categorical data. There are two possible reasons we saw what we saw in this experiment: since each indicator had at least four categories, the distributions was roughly normal but skewed because of the floor and ceiling effects; and since the sample sizes (or number of clusters) were less than 100 the WLS based estimators did not perform well. In fact, the performance did not began to get better until the sample size got to 100 provoking the question what would happen if the sample size was more. Research states you have to have at least five categories to threat the variable as normal but having at least four categories might have been the reason the ML based estimator performed well. The WLS based estimators require large sample sizes for continuous and categorical data according to Brown (2006), this could be the reason why we show the trend seen in all the charts. So, a sample size under 100 is not recommended when for WLS based estimators.

Limitations

The limitation of the study was threefold: sample size, categories, and replications. Although the number of replications was fine and the numbers were stable, higher number of reps is always good to ensure confidence. There were three sample sizes but there needs to be larger number sample sizes (and cluster sizes) so one can explore what happens when the cluster and/or sample size is very large. One can do this with a trend analysis so the selection of the cluster size and sample size is much more important if you want to test for a linear trend analysis (equal distance). Lastly, the number of categories does matter. It is known when there are five or more categories that you can treat the data as continuous. In this experiment, 3 thresholds were use to create 4 categories. Percentage of data I wanted for each category was dictated to Mplus. Mplus then tried to replicate the threshold for each indicator variable by making sure if I asked for 5% of the data to be from category 4 for a particular variable that I would obtain that amount within sampling variance (try to reproduce the threshold). Using three categories instead of four might produce different results. The reason ML based estimators performed better could have been because as the number of categories increase we began to treat the variable indicator as continuous.

Conclusion

As I said previously, WLS based estimators would have performed better if the sample size would have exceeded 100. However, there might be a golden ratio going on and as noted previously by Brown (2006) the sample size recommendation for WLSMV depended on the complexity of the model, so there are a number variable at play. For now, WLS based estimators are not recommended when the sample size is less than 100. For the ML based estimators a sample size of 30 is fine. Even though there were ceiling and floor effect, it seems there wasn’t enough excessive kurtosis to cause the sample size recommendation to increase.

APPENDIX C

Muthén's (2002) Code for CFA Model With Non-Normal Continuous Factor IndicatorsWithout Missing Data (*Generate two classes and analyze as one class*)

```
      TITLE: cfa3.inp non-normal, no missing
      MONTECARLO:          *Describes the monte carlo study
      NAMES ARE y1-y10;    *Names variables in data set
      NOBSERVATIONS = 265; *Specifies the total sample size
      NREPS = 10000;       *Specifies the number of replications
      SEED = 53487;        *User specified seed
      NCLASSES = 1;        *Now called classes, number of latent classes to
analyze
      GCLASSES = 2;        *Now genclasses, number of latent classes generate
      SAVE = cfa3.sav;
      ANALYSIS: TYPE = MIXTURE;   *Describes type of data mixture model
      ESTIMATOR = MLR;            *Describes Estimator used (Robust ML)

      MODEL MONTECARLO:           *Provide true model values use for
generation
      %OVERALL%
      f1 BY y1-y5*.8;             *gives the true model values for the
loadings
      f2 BY y6-y10*.8;
      f1@1 f2@1;                  *gives the factor variances
      y1-y5*.36 y6-y10*9;         *gives the error variances
      f1 WITH f2*.95;             *gives the covariances for factors one and
two
      [C#1@-2];                   *proportion of people in class one
(logit)12%
      %C#1%                       *class one [mean] and standard deviation
      [f1@0 f2@15];               *class one means, f1 mean and f2 mean
      f1@1 f2@5;                  *class one STD deviation
      %C#2%                       *class two mean and standard deviation
      [f1@0 f2@0];
      f1@1 f2@1;

      MODEL:                      *model to be estimated along with starting
values
      %OVERALL%
      f1 BY y1-y5*.8;
      f2 BY y6-y10*4;
      f1@1 f2@1;
      y1-y5*.36 y6-y10*9;
      f1 WITH f2*.20;
      [y6-y10*1.42];
      OUTPUT: TECH9;              *says to output the various error messages
```

APPENDIX D

Hox (2010) code for multilevel data (25 groups of size 10)

```
      Table A1. TITLE Simulation run for ML, ICC low, NG=25, GS=10;
      MONTECARLO:
      NAMES ARE y1-y6;
      NOBSERVATIONS = 250;    *total sample size, 25*10=250
      NREPS = 1000;
      SEED = 0;
      NCSIZES = 2;            *number of unique clusters
      CSIZES = 25 (3) 25 (7); *for unbalanced data; number of clusters and
sizes
      RESULTS = results01.sav;      *saves monte carlo results inan ASCII
file

      MODEL POPULATION:             *generates the model, true population
value
      %within%                      *the within level model values
      fw1 BY y1-y3@1;
      fw2 BY y4-y6@1;
      y1-y6@.10;
      fw1@.43;
      fw2@.43;
      fw1 WITH fw2@.172;
      %between%                     *the between level model values
      fb1 BY y1-y6@1;
      y1-y6@.25;
      fb1@1

      MODEL:                        *estimated model with starting values
      %within%
      fw1 BY y1-y3*1;
      fw2 BY y4-y6*1;
      y1-y6*.10;
      fw1@.43;
      fw2@.43;
      fw1 WITH fw2*.172;
      %between%
      fb1 BY y1-y6*1;
      y1-y6*.25;
      fb1@1;
      ANALYSIS:
      TYPE = TWOLEVEL;               *running a two level model
      ESTIMATOR = ML;
      TECH9;
```

REFERENCES

American Psychiatric Association. (2000). *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision.* Washington, DC: Author.

Apter, A., & King, R. A. (2006). Management of the depressed, suicidal child or adolescent. *Child and Adolescent Psychiatric Clinics of North America*, 15, 999-1013.

Asparouhov T., Muthén B (2010). Multiple imputation with Mplus. Mplus Web Notes

Bandalos, D. L. (2006). The use of Monte Carlo studies in structural equation modeling research. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 385-426). Greenwich, CT: Information Age Publishing.

Bandalos, D. L., & Leite, W. (2011). Use of Monte Carlo Studies in Structural Equation Modeling Research. *Unpublished manuscript*.

Barker, E.D., Arseneault, L., Brendgen, M., Fontaine, N., & Maughan, B. (2008). Joint development of bullying and victimization in adolescence: Relation to delinquency and self-harm. *Journal of the American Academy of Child and Adolescent Psychiatry, 47*, 1030–1038.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology, 51*(6), 1173-1182.

Batsche, G. M., & Knoff, H. M. (1994). Bullies and their victims: Understanding a pervasive problem in the schools. *School Psychology Review, 23*, 165-175.

Berthold, K.A., & Hoover, J. H. (2000). Correlates of bullying and victimization among intermediate students in the midwestern USA. *School Psychology International, 21,* 65-78.

Biddulph, S. (2003). *Alcohol: What's a Parent to Believe?* Center City, MN: Hazelden Publishing.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Branlinger, E. (1991). Social class distinctions in adolescents' reports of problems and punishment in school. *Behavioral Disorders, 17*(1), 36-46.

Brener, N. D., Billy, J. O. G., Grady, W.R. (2003). Assessment of factors affecting the validity of self-reported health-risk behavior among adolescents: Evidence from the scientific literature. *Journal of Adolescent Health, 33,* 436–57.

Brener, N. D., Kann, L, McManus, T., Kinchen, S. A., Sundberg, E. C., & Ross, J.G. (2002). Reliability of the 1999 Youth Risk Behavior Survey Questionnaire. *Journal of Adolescent Health, 31*, 336-342.

Brown, T. A. (2006). Confirmatory factor analysis for applied research. The Guilford Press.

Byrne, B. (2011). Structural Equation Modeling with Mplus: Basic concepts, applications, and programming (Multivariate Applications Series). Routledge Academic.

Cavazos-Rehg, P., Krauss, M., Spitznagel, E., Schootman, M., Cottler, L., & Bierut, L. (2010). Associations between multiple pregnancies and health risk behaviors among U.S. adolescents. *Journal of Adolescent Health, 37*(6), 600-603.

Centers for Disease Control and Prevention. Methodology of the Youth Risk Behavior Surveillance System, September 24, 2004. *Morbidity and Mortality Weekly Report, 53* (RR-12). Retrieved September 2012 from http://www.cdc.gov/mmwr/PDF/rr/rr5312.pdf

Centers for Disease Control and Prevention. Youth Risk Behavior Surveillance—United States, 2009, Surveillance Summaries, June 4, 2010. *Morbidity and Mortality Weekly Report, 59* (SS-5). Retrieved September 2012 from http://www.cdc.gov/mmwr/pdf/ss/ss5905.pdf

Chung, T., & Martin, C. S. (2002). Concurrent and discriminant validity of DSM-IV symptoms of impaired control over alcohol consumption in adolescents. *Alcoholism, clinical and experimental research, 26*(4), 485-92. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11981124

Chung, T., Martin, C. S., & Winters, K. C. (2005). Recent Developments in Alcoholism, 17(I), 5-27. doi: 10.1007/0-306-48626-1_1.

Cohen, J., & Geier, V. K. (2010). School climate research summary: January 2010. *School Climate Brief*, *1*(1). Retrieved September 2012 from http://www.schoolclimate.org/climate/documents/SCBrief_v1n1_Jan2010.pdf

Cottler, L. B., Schuckit, M. A., Helzer, J. E., Crowley, T., Woody, G., Nathan, P., & Hughes, J. (1995). The DSM-IV field trial for substance use disorders: major results. *Drug and alcohol dependence, 38*(1), 59-69; discussion 71-83. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7648998

De Leo, D., Burgis, S., Bertolote, J. M., et al. (2006). Definitions of suicidal behavior: Lessons learned from the WHO/EURO Multicentre Study. *Crisis*, *27*, 4-15.

De Leo, D., Burgis, S., Bertolote, J. M., Kerkhof, A. J., & Bille-Brahe, U. (2006). Definitions of suicidal behavior. *Crisis: The Journal of Crisis Intervention and Suicide Prevention*, *27*(1), 4-15.

Deas, D., Riggs, P., Langenbucher, J., Goldman, M., & Brown, S. (2000). Adolescents are not adults: developmental considerations in alcohol users. *Alcoholism, clinical and experimental research, 24*(2), 232-237. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10698377

DeWit, D. J., Adlaf, E. M., Offord, D. R., & Ogborne, A. C. (2000). Age at first alcohol use: a risk factor for the development of alcohol disorders. *American Journal of Psychiatry*, *157*(5), 745-750.

Dyer, N., Hanges, P., & Hall, R. (2005). Applying multilevel confirmatory factor analysis techniques to the study of leadership. *The Leadership Quarterly*, *16*(1), 149-167. doi:10.1016/j.leaqua.2004.09.009

Ellickson, P. L., & Hays, R. D. (1990). Beliefs about resistance self-efficacy and drug prevalence: do they really affect drug use? *Substance Use & Misuse*, *25*(S11), 1353-1378.

Ellickson, P. L., McGuigan, K. A., Adams, V., Bell, R. M., & Hays, R. D. (1996). Teenagers and alcohol misuse in the United States: by any definition, it's a big problem. *Addiction*, *91*(10), 1489-1503. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8917917

Enders, C. K. (2011). Missing not at random models for latent growth curve analyses. *Psychological methods*, *16*(1), 1.

Epstein, J. A., & Spirito, A. (2009). Risk factors for suicidality among a nationally representative sample of high school students. *Journal of Suicide and Life-Threatening Behavior, 39,* 241–251.

Evashwick, W. (1998). *Promoting teen health: Linking schools, health organizations, and community*. SAGE Publications, Incorporated.

Fabrigar, L. R., Porter, R. D., & Norris, M. E. (2010). Some things you should know about structural equation modeling but never thought to ask. *Journal of Consumer Psychology, 20,* 221–225.

Fan, X., Sivo, S., & Keenan, S. (2002). *SAS for Monte Carlo studies: A guide for quantitative researchers*. Sas Institute.

Finney, S. J., & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. *Structural equation modeling: A second course*, 269-314.

Fitzpatrick, K. M., Dulin, A., & Piko, B. (2010). Bullying and depressive symptomatology among low-income, African–American youth. *Journal of youth and adolescence*, *39*(6), 634-645.

Gilmartin, B. G. (1987). Peer group antecedents of severe love-shyness in males. *Journal of Personality, 55,* 467-489

Gladstone, G. L., Parker, G. B., & Malhi, G. S. (2006). Do bullied children become anxious and depressed adults? A cross-sectional investigation of the correlates of bullying and anxious depression. *Journal of Nervous Mental Disorders, 194*(3), 201–208.

Gold, K. F., & Muthén, B. (1991, April). Extensions of covariance structure analysis: hierarchical modeling of multidimensional achievement data. In *annual meeting of the American Educational Research Association, Chicago*.

Gottfredson, N. C., Panter, A. T., Daye, C. E., Allen, W. F., & Wightman, L. F. (2009). The effects of educational diversity in a national sample of law students : Fitting multilevel latent variable models in data with categorical indicators. *Diversity,* 305-331. doi:10.1080/00273170902949719

Geiser, C. (2013). Data analysis with Mplus. New York: Guilford Press.

Graham, J.W., Olchowski, A.E., Gilreath, T.D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science. 8(*3*),* 206-213.

Gray, J. (2010). *Why our drug laws have failed: a judicial indictment of war on drugs*. Temple University Press.

Green, S. B., Akey, T. M., Fleming, K. K., Hershberger, S. L., & Marquis, J. G. (1997). Effect of the number of scale points on chi-square fit indices in confirmatory factor analysis. *Structural Equation Modeling, 4,* 108-120.

Hacker, K. A., Suglia, S. F., Fried, L. E., Rappaport, N., & Cabral, H. (2006). Developmental differences in risk factors for suicide attempts between ninth and eleventh graders. *Suicide and Life-Threatening Behavior*, *36*(2), 154-166.

Hacker, K. A., Suglia, S. F., Fried, L. E., Rappaport, N., & Cabral, H. (2006). Developmental differences in risk factors for suicide attempts between ninth and eleventh graders. *Suicide & Life-Threatening Behavior, 36*(2), 154-166. doi:10.1521/suli.2006.36.2.154

Haynie D. L., Nansel T., Eitel P., Crump, A. D., Saylor, K., Yu K., & Morton, B. (2001). Bullies, victims, and bully-victims: Distinct groups of at-risk youth. *Journal of Early Adolescence 21*, 29-49.

Hays, R. D., & Ellickson, P. L. (1996). What is adolescent alcohol misuse in the United States according to the experts? *Alcohol and alcoholism, 31*(3), 297-303. Retrieved September 2012 from http://www.ncbi.nlm.nih.gov/pubmed/8844036

Hazler, R. J. (1992). What kids say about bullying. *Executive Educator*, *14*(11), 20-22.

Hendin, H., Maltsberger, J. T., Lipschitz, A., Haas, A. P., & Kyle, J. (2001). Recognizing and responding to a suicide crisis. *Annals of the New York Academy of Sciences, 932*(1), 169-187.

Hendin, H., Maltsberger, J. T., Lipschitz, A., Haas, A. P., & Kyle, J. (2001). Recognizing and responding to a suicide crisis. *Journal of Suicide and Life-Threatening Behavior, 31*, 115–128.

Hinduja, S., & Patchin, J. W. (2010). Bullying, cyberbullying, and suicide. *Archives of Suicide Research, 14*(3), 206-221.

Horne, A. M., Newman-Carlson, D., & Bartolomucci, C. L. (2003). *Bully Busters: A Teacher's Manual for Helping Bullies, Victims, and Bystanders: Grades K-5*. Research PressPub.

Hox, J. J. (1995). *Applied Multilevel Analysis.* Amsterdam, The Netherlands: TT-Publikaties.

Hox, J. J. (2002). Multilevel analysis: Techniques and applications. Mahwah, NJ: Lawrence Erlbaum Associates.

Hox, J. J., Maas, C. J. M., & Brinkhuis, M. J. S. (2010). The effect of estimation method and sample size in multilevel structural equation modeling. *Statistica Neerlandica, 64*, 157-170.

Hox, J. J., Maas, C. J. M., & Brinkhuis, M. J. S. (2010). The effect of estimation method and sample size in multilevel structural equation modeling. *Statistica Neerlandica, 64*(2), 157-170. doi:10.1111/j.1467-9574.2009.00445.x

http://cranepsych.edublogs.org/files/2009/07/dark_zero_tolerance.pdf

Hu, L.-T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3,* 424-453.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal, 6(1), 1-55.

Iacobucci, D. (2008). *Mediation analysis* (No. 156). SAGE Publications, Incorporated.

Ivanoff, A. (1989). Identifying psychological correlates of suicidal behavior in jail and detention facilities. *Psychiatric Quarterly, 60,* 73-84.

Ivarsson, T., Broberg, A. G., Arvidsson, T., & Gillberg, C. (2005). Bullying in adolescence: Psychiatric problems in victims and bullies as measured by the Youth Self Report (YSR) and the Depression Self-Rating Scale (DSRS). *Nordic Journal of Psychiatry, 59*(5), 365-373.

Ivarsson, T., Broberg, A. G., Arvidsson, T., & Gillberg, C. (2005). Bullying in adolescence: psychiatric problems in victims and bullies as measured by the Youth Self Report (YSR) and the Depression Self-Rating Scale (DSRS). *Nordic Journal of Psychiatry, 59*(5), 365-373. doi:10.1080/08039480500227816

Jiang, Y., Perry, D., & Hesser, J. (2010). Adolescent suicide and health risk behaviors: Rhode Island’s 2007 Youth Risk Behavior Survey. *American Journal of Preventative Medicine, 38*(5), 551-555.

Joreskog, K. G., & Sorbom, D. (1982). Recent developments in structural equation modeling. *Journal of Marketing Research, 19,* 404-416

Julian, M. W. (2001). The consequences of ignoring multilevel data structures in nonhierarchical covariance modeling. *Structural Equation Modeling*, *8*(3), 325-352.

Kalman, I. (2001, August 11). Norway Massacre: What It Reveals About the Olweus Bullying Program, Is Norwegian compassion the secret to Olweus’ positive research results? Retrieved September, 2012, from http://bullies2buddies.com/Articles/norway-massacre-what-it-reveals-about-the-olweus-bullying-program.html.

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling*, *15*(1), 136-153.

Kamata, A., & Bauer, D. J. (2008). A Note on the Relation between Factor Analytic and Item Response Theory Models. *Structural Equation Modeling: A Multidisciplinary Journal, 15*(1), 136-153. doi:10.1080/10705510701758406

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook*. Upper Saddle River, N.J: Pearson Prentice Hall.

Kim, Y. S., & Leventhal, B. (2008). Bullying and suicide: A review. *International Journal of Adolescent Medical Health, 20*(2), 133-154.

Klangphahol, K., Traiwichitkhun, D., & Kanchanawasi, S. (n.d.). Applying multilevel confirmatory factor analysis techniques to perceived homework quality.

Klomek, A. B., Kleinman, M., Altschuler, E., Marrocco, F., Amakawa, L., & Gould, M. S. (2011). High school bullying as a risk for later depression and suicidality. *Suicide and Life Threatening Behavior, 41*(5), 501-516.

Klomek, A. B., Marrocco, F., Kleinman, M., Schonfeld, I. S., & Gould, M. S. (2007). Bullying, depression, and suicidality in adolescents. *Journal American Academy of Child and Adolescent Psychiatry, 46*(1), 40.

Klomek, A. B., Marrocco, F., Kleinman, M., Schonfeld, I.S., & Gould, M. S. (2008). Peer victimization, depression, and suicidiality in adolescents. *Suicide and Life-Threatening Behavior, 38,* 166-180.

Klomek, A. B., Sourander, A., & Gould, M. S. (2011). Bullying and suicide. *Psychiatric Times, 28*(2), 27-31.

Kupek, E. (2006). Beyond logistic regression: structural equations modelling for binary variables and its application to investigating unobserved confounders. *BMC Medical Research Methodology*, *6*(1), 13.

Larkin, R. W. (2007). *Comprehending Columbine*. Temple University Press.

Lei, M., & Lomax, R. G. (2005). The effect of varying degrees of nonnormality in structural equation modeling. *Structural Equation Modeling, 12*, 1-27.

Lester, P. E., & Bishop, L. K. (2000). *Handbook of tests and measurement in education and the social sciences*. Rowman & Littlefield Education.

Lester, P. E., & Bishop, L. K. (2000). *Handbook of tests and measurement in education and the social sciences*. Rowman & Littlefield Education.

Li, F., Duncan, T. E., Harmer, P., Acock, A., & Stoolmiller, M. (1998). Analyzing measurement models of latent variables through multilevel confirmatory factor analysis and hierarchical linear modeling approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, *5*(3), 294-306.

Liang, H., Flisher, A. J., & Lombard, C. J. (2007). Bullying, violence, and risk behavior in South African school students. *Child Abuse and Neglect, 31,* 161-171

Maas, C. J. M, & Hox, J. J. (2005). Sufficient Sample Sizes for Multilevel Modeling. Methodology: *European Journal of Research Methods for the Behavioral and Social Sciences*, *1*(3), 86-92. doi:10.1027/1614-1881.1.3.86

Maas, C. J., & Hox, J. J. (2002). Robustness of multilevel parameter estimates against small sample sizes. *Unpublished Paper. The Netherlands: Utrecht University*.

Maas, C.J.M. & Hox, J.J. (2001). Sample sizes for multilevel modeling. Retrieved October, 2012 from http://www.upa.pdx.edu/IOA/newsom/semclass/ho_estimate2.pdf

MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research*, *39*(1), 99-128.

Maney, D. W., Higham-Gardill, D. A., & Mahoney, B. S. (2002). The alcohol-related psychosocial and behavioral risks of a nationally representative sample of adolescents. *Journal of School Health*, *72*(4), 157-163.

Maney, D. W., Higham-Gardill, D. A., & Mahoney, B. S. (2002). The alcohol-related psychosocial and behavioral risks of a nationally representative sample of adolescents. *Journal of School Health*, *72*(4), 157-163.

Martins, S. S., & Alexandre, P. K. (2009). The association of ecstasy use and academic achievement among adolescents in two US national surveys. *Addictive behaviors*, *34*(1), 9-16.

Mason, W. A., Hitch, J. E., Kosterman, R., Mccarty, C. A., Herrenkohl, T. I., & Hawkins, J. D. (2010). Growth in adolescent delinquency and alcohol use in relation to young adult crime , alcohol use disorders , and risky sex : A comparison of youth from low- versus middle-income backgrounds. *Journal of Child Psychology and Psychiatry, 12*, 1377-1385. doi:10.1111/j.1469-7610.2010.02292.x

Mason, W.A., Hitch, J. E., Kosterman, R., McCarty, C. A., Herrenkohl, T. I., & David Hawkins, J. (2010). Growth in adolescent delinquency and alcohol use in relation to young adult crime, alcohol use disorders, and risky sex: a comparison of youth from low-versus middle-income backgrounds. *Journal of child psychology and psychiatry*, *51*(12), 1377-1385.

May, A., & Klonsky, E. D. (2011). Validity of suicidality items from the Youth Risk Behavior Survey in a high school sample. *Assessment*, *18*(3), 379-381.

Mayer, M. J., & Leone, P. E. (1999). A Structural Analysis of School Violence and Disruption : Implications for Creating Safer Schools. *Children, 22*(3), 333-356.

Mayer, M. J., & Leone, P. E. (1999). A structural analysis of school violence and disruption: Implications for creating safer schools. *Education and Treatment of Children, 22*(3), 333–356.

Mayo, D. J. (1992). What is being predicted? The definition of “suicide.” In R. W. Maris, A. L. Berman, J. T. Maltsberger, & R. I. Yufit (Eds.), *Assessment and prediction of suicide* (pp. 88-101). New York: Guilford.

Mays, D., & Thompson, N. J. (2009). Alcohol-related risk behaviors and sports participation among adolescents: An analysis of 2005 Youth Risk Behavior Survey Data. *Journal of Adolescent Health, 44*(1), 87-89.

McNeely, C. A., Nonnemaker, J. M., & Blum, R. W. (2002). Promoting school connectedness: Evidence from the national longitudinal study of adolescent health. *Journal of School Health*, *72*(4), 138-146.

Meuleman, B., & Billiet, J. (2009). A Monte Carlo sample size study : How many countries are needed for accurate multilevel SEM ? *Methods, 3*(1), 45-58.

Meuleman, B., & Billiet, J. (2009). A Monte Carlo sample size study: How many countries are needed for accurate multilevel SEM? In *Survey research methods* (Vol. 3, No. 1, pp. 45-58).

Miller, J. W., Naimi, T. S., Brewer, R. D., & Jones, S. E. (2007). Binge drinking and associated health risk behaviors among high school students. *Pediatrics*, *119*(1), 76-85.

Moineddin, R., Matheson, F. I., & Glazier, R. H. (2007). A simulation study of sample size for multilevel logistic regression models. *BMC Medical Research Methodology*, *7*(1), 34.

Muthén, B. (2002). Using Mplus Monte Carlo simulations in practice: A note on non-normal missing data in latent variable models. *Mplus webnotes*, (2).

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement*, *28*(4), 338-354.

Muthén, B. O. (1994). Multilevel Covariance Structure Analysis. *Sociological Methods & Research*, *22*(3), 376-398. doi:10.1177/0049124194022003006

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological methods & research*, *22*(3), 376-398.

Muthén, L.K. and Muthén, B. O. (1998-2010). Mplus User's Guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén

Muthén, B., & Asparouhov, T. (2002). Using Mplus Monte Carlo simulations in practice: A note on non-normal missing data in latent variable models. *Mplus webnotes*, (2).

Muthén, L. K., & Muthén, B. (2009). Multilevel modeling with latent variables using Mplus: cross sectional analysis. *Unpublished manuscript*, *Berlin*.

Muthén, L. K., & Muthén, B. (February 2012). Two Factor CFA Example in Mplus. Retrieved October 2012 from http://www.upa.pdx.edu/IOA/newsom/semclass/ho_cfa2.pdf

Nansel, T. R., Craig, W., Overpeck, M. D., Saluja, G., & Ruan, W. (2004). Cross-national consistency in the relationship between bullying behaviors and psychosocial adjustment. *Archives of Pediatrics & Adolescent Medicine*, *158*(8), 730.

Nansel, T. R., Overpeck, M. D., Haynie, D. L., Ruan, W., & Scheidt, P. C. (2003). Relationships between bullying and violence among US youth. *Archives of Pediatrics & Adolescent Medicine*, *157*(4), 348.

National School Climate Center. (2012). School climate. Retrieved from http://www.schoolclimate.org/climate/

Newman-Carlson, D., Horne, A. M., & Bartolomucci, C. L. (2000). *Bully Busters: A Teacher's Manual for Helping Bullies, Victims, and Bystanders:[Grades 6-8]*. Research Press.

Newsom, J. (2010). Practical approaches to dealing with nonnormal and categorical variables. Retrieved October 22, 2012, from http://www.upa.pdx.edu/IOA/newsom/semclass/ho_estimate2.pdf

O'Carroll, P. W., Berman, A. L., Maris, R. W., Moscicki, E. K., Tanney, B. L., & Silverman, M. M. (1996). Beyond the Tower of Babel: a nomenclature for suicidology. *Suicide and Life-Threatening Behavior*, *26*(3), 237-252.

Olweus Bullying Prevention Program. (2011). Retrieved September, 2012, from http://www.clemson.edu/olweus/

Olweus D. (1992). Bullying among schoolchildren: Intervention and prevention. In: R.D. Peters, R.J. McMahon, V.L. Quinsey, eds. Aggression and Violence Throughout the Life Span. London, England: Sage Publications, 100-125.

Olweus, D. (1994). Bullying at school: basic facts and effects of a school based intervention program. *Journal of child psychology and psychiatry*, *35*(7), 1171-1190.

Olweus, D. (1997). Bully/victim problems in school: Facts and intervention. *European Journal of Psychology of Education*, *12*(4), 495-510.

Olweus, D., & Alsaker, F. D. (1991). Assessing change in a cohort-longitudinal study with hierarchical data. *Problems and methods in longitudinal research: Stability and change*, 107-132.

Olweus. D & Limber, S.P. (2012). The Olweus Bullying Prevention Program: Implementation and Evaluation over Two Decades. *Unpublished manuscript*. Retrieved September 2012 from http://www.bullyingpreventioninstitute.org/LinkClick.aspx?fileticket=5BnCPJGFPhc%3D&tabid=72

Orpinas, P., & Horne, A. M. (2005). *Bullying prevention: Creating a positive school climate and developing social competence*. APA Books. Available from: American Psychological Association, 750 First Street NE, Washington, DC 20002.

Osterman, K. F. (2000). Students' need for belonging in the school community. *Review of Educational Research*, *70*, 323–367.

Paxton, R.J., Valois, R.F., Watkins, K.W., Huebner, S., Wazner Drane J. (2007). Associations between depressed mood and clusters of health risk behaviors. *American Journal of Health Behavior*, *31*, 272–283.

Peskin, M. F., Tortolero, S. R., Markham, C. M., Addy, R. C., & Baumler, E. R. (2007). Bullying and victimization and internalizing symptoms among low-income black and hispanic students. *Journal of Adolescent Health*, *40*(4), 372-375.

Pintado, I. (2006). *Perceptions of school climate and bullying in middle schools* (Doctoral dissertation, University of South Florida). Retrieved September, 2012, from http://digital.lib.usf.edu:8080/fedora/get/usfldc:E14-

SFE0001816/DOCUMENThttp://digital.lib.usf.edu:8080/fedora/get/usfldc:E14-SFE0001816/DOCUMENT

Preacher, K. J., Zhang, Z., & Zyphur, M. J. (2011). Alternative methods for assessing mediation in multilevel data: The advantages of multilevel SEM. *Structural Equation Modeling*, *18*(2), 161-182.

Preacher, K. J., & Selig, J. P. (2012). Advantages of Monte Carlo confidence intervals for indirect effects. Communication Methods and Measures, 6(2), 77-98.

Raykov, T., and Marcoulides, G. A. 2006. *A First Course in Structural Equation Modeling* (2nd ed.), Mahwah, NJ: Lawrence Erlbaum Associates.

Reboussin, B. A., Song, E. Y., Shrestha, A., Lohman, K. K., & Wolfson, M. (2006). A latent class analysis of underage problem drinking: Evidence from a community sample of 16–20 year olds. *Drug and alcohol dependence*, *83*(3), 199.

Reise, S. P., Ventura, J., Nuechterlein, K. H., & Kim, K. H. (2005). An illustration of multilevel factor analysis. *Journal of personality assessment*, *84*(2), 126-136.

Retrieved September, 2012, from http://www.indiana.edu/~safeschl/ztze.pdf

Rigby, K. (1999). Peer victimization at school and the health of secondary students. *British Journal of Educational Psychology*, *69 (1)*, 95-104.

Roesch, S. C., Aldridge, A. A., Stocking, N., Villodas, F., & Leung, Q., Bartley, C.E., & Black, L.J. (2010). Multilevel Factor Analysis and Structural Equation Modeling of Daily Diary Coping Data : Modeling Trait and State Variation, *Multivariate Behavioral*, *45(5)*, 767-789.

Rucker, D. D., Preacher, K. J., Tormala, Z. L., & Petty, R. E. (2011). Mediation analysis in social psychology: Current practices and new recommendations. *Social and Personality Psychology Compass*, *5*(6), 359-371.

Rudd, M. D., Berman, A. L., Joiner, T. E., Nock, M. K., Silverman, M. M., Mandrusiak, M., ... & Witte, T. (2006). Warning signs for suicide: Theory, research, and clinical applications. *Suicide and Life-Threatening Behavior*, *36*(3), 255-262.

Sampson, R. (2002). Bullying in Schools: Problem-Oriented Guides for Police Series. Retrieved September, 2012, from http://www.cops.usdoj.gov/pdf/e12011405.pdf

Satorra, A., & Muthén, B. (1995). Complex sample data in structural equation modeling. *Sociological methodology*, *25*, 267-316.

Schafer, J.L., & Graham, J.W. (2002). Missing data: Our view of the state of the art. Psychological Methods, 7, 147-177.

Schilling, E. A., Aseltine Jr, R. H., Glanovsky, J. L., James, A., & Jacobs, D. (2009). Adolescent alcohol use, suicidal ideation, and suicide attempts. *Journal of Adolescent Health*, *44*(4), 335-341.

Schumaker, R. E., & Lomax, R. G. (2004). *A beginner's guide to structural equation modeling*

Seals, D., & Young, J. (2003). Bullying and victimization: prevalence and relationship to gender, grade level, ethnicity, self-esteem, and depression. *Adolescence*, *38*(152), 735.

Shneidman, E. S. (1998). *The suicidal mind*. Oxford University Press, USA.

Silverman, M. M., & Maris, R. W. (1995). The prevention of suicidal behaviors: An overview. *Suicide and Life-Threatening Behavior, 25*, 10-21.

Skiba, R. J. (2000). Zero tolerance, zero evidence. *An analysis of school disciplinary practice*.

Skiba, R., & Peterson, R. (1999). The dark side of zero tolerance: Can punishment lead to safe schools?. *The Phi Delta Kappan*, *80*(5), 372-382. Retrieved on September, 2012 from

Skrondal, A., & Rabe-Hesketh, S. (2008). Multilevel and related models for longitudinal data. In *Handbook of multilevel analysis* (pp. 275-299). Springer New York.

Spearman, C. (1904). The proof and measurement of association between two things. *The American journal of psychology*, *15*(1), 72-101.

Stapleton, L. M. (2006). An assessment of practical solutions for structural equation modeling with complex sample data. *Structural Equation Modeling*, *13*(1), 28-58.

Stapleton, L. M. (2006). Using multilevel structural equation modeling techniques with complex sample data. In G. R. Hancock & R. Mueller (Eds.), *Structural equation modeling: A second course*, Greenwich, CT: Information Age Publishing, 345-383.

Stein, J. A., Dukes, R. L., & Warren, J. I. (2007). Adolescent male bullies, victims, and bully-victims: A comparison of psychosocial and behavioral characteristics. *Journal of pediatric psychology*, *32*(3), 273-282.

Substance Abuse and Mental Health Services Administration. (2007). Results from the 2006 National Survey on Drug Use and Health: National Findings (Office of Applied Studies, NSDUH Series H-32, DHHS Publication No. SMA 07-4293). Rockville, MD. Retrieved September, 2012, from http://oas.samhsa.gov/NSDUH/2k6NSDUH/2k6results.cfm#Ch3

Substance Abuse and Mental Health Services Administration. (2010). Results from the 2009 National Survey on Drug Use and Health: Volume I. Summary of National Findings (Office of Applied Studies, NSDUH Series H-38A, HHS Publication No. SMA 10-4856Findings). Rockville, MD.

Tabachnick, B. G., & Fidell, L. S. (2007). Experimental designs using ANOVA. Australia: Thomson/Brooks/Cole.

Twisk, J. W. (2006). *Applied multilevel analysis: a practical guide*. Cambridge University Press.

Weinberg, S. L., & Abramowitz, S. K. (2002). Data analysis for the behavioral sciences using SPSS. Cambridge, UK: Cambridge University Press.

Wright, S. (1932, January). The roles of mutation, inbreeding, crossbreeding and selection in evolution. In *Proceedings of the sixth international congress on genetics* (Vol. 1, No. 6, pp. 356-366).

Wu, S. C., Pink, W., Crain, R., & Moles, O. (1982). Student suspension: A critical reappraisal. *The Urban Review*, *14*(4), 245-303.

Yu, C. Y., & Muthén, B. (2002, April). Evaluation of model fit indices for latent variable models with categorical and continuous outcomes. In *annual meeting of the American Educational Research Association, New Orleans, LA*.

Zhao, X., Lynch, J. G., & Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and truths about mediation analysis. *Journal of Consumer Research*, *37*(2), 197-206.