A Consideration of Issues Related to the Confirmatory Factor Analysis of the Evaluation Use and Evaluation Involvement Scales

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Gina Marie Johnson

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Frances Lawrenz, Adviser

May, 2011

Acknowledgements

It has been a phenomenal learning experience completing the requirements for a PhD at the University of Minnesota. I received tremendous support from family and friends and treasure the relationships I have developed during these years of study. There are a number of people who deserve particular acknowledgement.

My adviser and mentor, Dr. Frances Lawrenz

Thank you for everything. Really. I can't possibly list it all. You taught me so much about the process of completing research using both quantitative and qualitative methods. I can never hope to match your work ethic, but consider it the highest of compliments every time a colleague notes the similarities between us.

My committee members, Drs. Bob delMas, Jean King, and Michael Rodriguez

Bob, you are such a gifted teacher of statistics and provided for me an example of how to successfully combine research with practice.

Jean, you are so wonderfully skilled at working with people and I hope to emulate your combination of intelligence, kindness, and enthusiasm for educational evaluation.

Michael, you display a sense of thoughtful consideration I hope to imitate as I use the measurement and survey design skills you taught me to conduct educational research.

My classmates

There are so many wonderful people with whom I studied and from whom I learned so much. In particular I must thank Denise Roseland, who patiently caught me up on everything related to the Evaluation Use Grant and answered my many questions about the PhD process; Stacy Karl and Julio Cabrera, with whom I spent countless hours studying, completing coursework, and successfully completing some challenging courses; and Chris Desjardins, who was the best officemate a person could ever ask for and showed me that I could actually learn to use R.

My sisters, Kim Adler and Heidi Danielson

I have such fond memories of playing student to your teachers at the antique desk in the basement on Hull Road. It is likely that I am one of the only children to have entered kindergarten knowing the multiplication tables up to 5's and with an understanding of the Vikings' colonization of Greenland. You paved the way for me to college and beyond, and for that I will always be grateful.

My parents, Joe and Marilyn Adler

You instilled in me the importance of education from the beginning and I cannot thank you enough for the support you gave me as a child and continue to provide me as an adult. If I received nothing more from these years of study than your praise, it would all be worth it, but luckily the rewards are far greater. Thank you for taking an interest in my studies even when they became so specialized as to be downright boring. Your questions about my progress were evidence of your pride in my work.

My best friend and husband, Sean Johnson

You said from the beginning that I could do this and that you would support me all the way. You were my Sherpa on this, my version of Mount Everest. Thank you for calming me down and helping me find answers when I was tired, frustrated, or down on myself. Thank you for noticing that taking four difficult classes in one semester brought out the best in my skills and work ethic, and for telling me that you noticed because it made the hard work more worthwhile. Now it's your turn to be supported by me for awhile, so go stock up on coconuts to slice open and sell at the beach.

Dedication

This dissertation is dedicated to my two grandmothers:

Alberta Laufenberg, who taught me the importance of hard work through her example.
Leona Adler, who taught me to cherish education through her desire to attend high school.

It was hard, Grandma Laufenberg, but I persevered.
You couldn't go to school beyond 8th grade, Grandma Adler, so I earned a PhD.

Abstract

Factor analysis as a methodological technique has been continually improved and updated for use with data with a variety of characteristics, though the default settings on most software packages assume the use of continuous, normally distributed data. Evaluators planning to use confirmatory factor analysis (CFA) with ordinal survey data measured with intensity response format items must be aware of the special characteristics of their data as well as the decisions these characteristics necessitate in the CFA process. While completing CFAs on data collected using the Evaluation Use and Evaluation Involvement Scales, data characteristics and analysis criteria specific to ordinal data were considered. Guidelines for evaluators interested in completing CFA of data with similar characteristics are presented.

Table of Contents

List of Tables

List of Figures

CHAPTER 1 - INTRODUCTION

> *If the misapplication of factor methods continue at the present rate, we shall soon find general disappointment with the results because they are usually meaningless as far as psychological interpretation is concerned* (Thurstone, 1937, p. 73).

> *Americans today are accustomed to a seemingly endless stream of questions from survey researchers, political pollsters, marketers, and census takers… Being studied… is an understood and unexceptional feature of modern life* (Igo, 2007, p. 3).

In the early days of the exploration of factor analysis as a measurement tool in the field of psychology, Louis Thurstone warned of the potential for misuse of this now commonplace analysis procedure. Seventy years later, in 2007, Sarah Igo paints a picture of a society where surveys are ubiquitous and people's opinions are measured constantly. For evaluators and researchers, it is critical for measurement tools to be viewed as showing both reliability, defined as precision and accuracy, and validity, defined as relevancy to the inferences being made with the measurement tool (Thorndike, 2005). The process by which the creators of these tools, such as surveys designed for research or evaluation, provide evidence to support their use for the intended purpose, can involve a series of expert review, pilot testing, survey administration, and statistical analyses. Factor analysis is just one potential tool in the measurement instrument development toolkit available to survey developers. But, as Thurstone warned in 1937, it is a tool that can be misapplied, leaving the developer with questionable results that may negatively impact the believability of

results obtained from the survey. It is critical for those who choose to use surveys in their research and evaluation work to apply factor analysis correctly.

Confirmatory factor analysis and survey development

Introduced as a concept by Spearman in 1904 and further refined by Thurstone in the 1930's and beyond, factor analysis involves partitioning of the variance of each indicator into common variance (accounted for by the latent factor) and unique variance (Brown, 2006). Currently, factor analysis is considered to be the most commonly used procedure in developing latent variable measures (Floyd & Widaman, 1995). Exploratory factor analysis (EFA) is often used in situations where data reduction is a main goal, such as lengthy surveys in which an evaluator or researcher wishes to summarize a respondent's answers to multiple questions in one data point (Brown). Due to its exploratory nature, it is also often used in the early stages of instrument development, when the survey's creator is working to better understand how the items on a survey, or factors of a latent variable, relate to one another (Floyd & Widaman). Confirmatory factor analysis (CFA) is more often used to test a priori hypotheses held by the evaluator or researcher creating the measurement instrument. CFA can be used to provide validity evidence for the use of a survey and also to compare the fit of different latent construct models under consideration by the researcher or evaluator (Floyd & Widaman).

Surveys used in research and evaluation often measure attitudes, opinions, and amounts using what is often referred to as a "Likert scale" or Likert-

type scale". Due to the confusion in the literature on the meaning of the term scale (see Carifio & Perla, 2007), and consequently the properties related to statistical analyses of surveys that utilize this form of measurement, a researcher interested in survey scale development must first sort out the variety of definitions to the terms associated with surveys. A more descriptive term, such as intensity response format, might be more accurate and clear up some of the confusion in the literature. As a researcher completes the process of survey development, including item writing, pilot testing, data collection, and studies of reliability and validity for the instrument, she must consider how best to measure the construct of interest and how that measurement decision will impact the later procedures inherent in this development process. CFA may need to be conducted in a particular manner when the data used for the procedure are measured with intensity response format items.

Issues related to the use of ordinal data in surveys and CFA

In 1946, a committee of the British Association for the Advancement of Science attempted to clarify the types of statistical analyses that can safely be conducted on different types of variables (Stevens). More contemporary researchers do not, however, universally agree with the historical committee's decisions (Knapp, 1990). Along with the variations on usage of the term scale in the literature on survey development, there also exists confusion in terms of the rules for determining whether a particular item or scale on a survey should be considered an ordinal variable, as well as what types of statistical analyses are

therefore appropriate to conduct on data gathered with those items and scales (Baker, Hardyck, & Petrinovich, 1966; Labovitz, 1967; O'Brien, 1979; Marcus-Roberts & Roberts, 1987; Knapp; Jamieson, 2004; Carifio & Perla, 2007). Ordinal variables cannot, by definition, be considered normally distributed, and normal distribution is a necessary assumption of many statistical analyses (Harwell & Gatti, 2001). CFA generally requires multivariate normality (Brown, 2006). Using ordinal data to conduct a CFA, then, should cause a researcher concern that she might, in fact, be misapplying factor analysis, as Thurstone cautioned. Maximum likelihood estimation, the most common estimation method used in CFA, requires data to be multivariate normally distributed and continuous, which are characteristics that cannot be attributed to ordinal data (Flora & Curran, 2004).

Solutions to the issues related to CFA and ordinal data

While a straightforward CFA using the default settings in a software program designed to complete the analysis would be the easiest way to perform a CFA in the process of developing a survey instrument, it would not necessarily be the correct way to complete the CFA and might lead to questionable results. Researchers in the field of psychology and measurement have been studying the issues related to CFA and ordinal data for the past few decades and have agreed on some potential solutions. Issues of particular concern to these researchers include methods for estimating parameters, measures of correlation, and fit indices. Other issues relevant to proper completion of CFA with ordinal data

include sample size, model size, number of indicators per factor, and number of categories per item. These issues can also interact in ways that cause further concerns for the evaluator or researcher developing a survey.

Regarding the correct way to estimate parameters in CFA with ordinal data, there seems to be general agreement that weighted least squares is the best method to use (Jöreskog & Moustaki, 2001; DiStefano, 2002; Flora & Curran, 2004). However, Muthén and Kaplan (1985) provide a slightly different opinion. Their study was conducted in the 1980s and was one of the early studies in this topic, so it appears that later researchers refined the research and came to general agreement. Polychoric correlation is the generally agreed upon method for measuring correlations in CFA with ordinal data (DiStefano, 2002; Flora & Curran, 2004; Babakus, Ferguson, & Jöreskog, 1987), though Babakus, Ferguson, and Jöreskog have several cautions for the researcher utilizing this technique.

Hutchinson & Olmos (1998) offer a summary review of a number of other studies that guided them toward a conclusion regarding the proper indices to use to measure fit. However, it appears that more research could be done in this area to provide more definitive conclusions. Because it is difficult to complete Monte Carlo simulations with every possible combination of characteristics that might occur in the real world when an evaluator or researcher is completing a CFA as part of the survey development process, there are not hard and fast rules related to the many combinations of characteristics of the data and model one

uses in a CFA. Muthén and Kaplan (1985) provide guidance related to minimum sample size needed for a CFA's results to be considered valid. Dolan (1994) offers suggestions about the relationship between sample size and the number of categories per item used in the survey, which is a decision survey developers make early in the process of designing the instrument. Finally, Potthast (1993) cautions those conducting CFA to consider the relationship between the available sample size and the number of indicators per factor in the latent variable model under consideration.

Evaluation Use and Evaluation Involvement Scales

From 2005 to 2009, a team of researchers at the University of Minnesota studied use and involvement in large-scale, multi-site evaluation studies funded by the National Science Foundation (NSF). During the course of their research, the team discovered that there is a lack of, and need for, instruments with which to measure both use of an evaluation and involvement in an evaluation process (Toal, et. al., 2006). Using an extensive review of the theoretical background of both evaluation use and evaluation involvement, as well as many other research tools, the team designed two sets of scales to fill this need in the field. The Evaluation Use and Evaluation Involvement Scales underwent EFA in the process of developing the survey the team administered to persons involved in evaluations of programs funded by the NSF. In order to provide construct validity evidence for the scales, it was determined that CFAs should be conducted on the scales after a new set of respondents completed the survey in which the scales

were included. Given the previous discussion of the issues related to the CFA of ordinal data, however, concern was expressed over the potential for questionable results if the procedure was completed incorrectly or if the data did not posess the characteristics necessary to meet the assumptions of the analyses. This study was designed to answer those concerns, with an in-depth consideration of CFA of two sets of scales using data with very specific characteristics.

Research Question

The question this study intends to address is: *What are the data characteristics and analysis criteria that need to be considered to meet the assumptions of confirmatory factor analysis of ordinal evaluation survey data?* Given the fact that the field of evaluation is in need of measurement tools with which to assess both use of and involvement in evaluation and the fact that surveys are used extensively when conducting evaluations in the social sciences (DiStefano, 2002) it is therefore critical for evaluators who use CFA in the process of survey development to understand the proper way to complete the analysis. It is also important to understand the data characteristics necessary for CFA completion and to compare these to what is realistically feasible in real-life evaluation situations. This study makes these considerations in the context of providing validity evidence for the Evaluation Use and Evaluation Involvement Scales. Lessons learned in the process will be presented as a set of guidelines for others who wish to conduct CFAs on data with similar characteristics.

CHAPTER 2 - LITERATURE REVIEW

Introduction

Some confusion seems to exist in the literature regarding the use of ordinal items and scales in surveys and measurement (Baker, Hardyck, & Petrinovich, 1966; Labovitz, 1967; O'Brien, 1979; Marcus-Roberts & Roberts, 1987; Knapp, 1990; Jamieson, 2004; Carifio & Perla, 2007). Much of the disagreement stems from a lack of consistency in definition of terms, particularly the term "scale" (Carifio & Perla). This definitional confusion also exists as statistical confusion in that a clear set of recommendations for completing CFA on ordinal data cannot yet be clearly summarized from the simulation studies that have been conducted. This literature review is designed to guide the reader through the maze of literature on ordinal survey items and scales and subsequent CFA of these ordinal variables in order to summarize the findings up to the present time.

Because this project began with a desire to properly complete confirmatory factor analyses of two sets of scales that had previously undergone exploratory factor analyses, the review of the literature will begin with an overview of the theoretical frameworks behind the creation of the Evaluation Use and Evaluation Involvement Scales, which are the two sets of scales in question. This overview will allow the reader to better understand the theory behind the development of the two sets of scales and provide evidence for the need for

construct validity evidence for the scales, which the CFAs could potentially provide. Following this overview, a summary of the literature describing the history of factor analysis as a data analytic technique in developing surveys and providing validity evidence for their use will be presented. Finally, a review of the literature relevant to the disagreement surrounding ordinal variables in survey items, scales, and CFA will be summarized, including clarification of the terms used by this researcher as definitional issues are a major part of the controversy. This presentation of the literature relevant to the three sections, the Evaluation Use and Evaluation Involvement Scales, the history and use of factor analysis, and the confusion surrounding analysis of ordinal data, will leave the reader with an understanding of the need for further research in this area.

The Evaluation Use and Evaluation Involvement Scales

Evaluation Use and Involvement Defined

Before elaborating on the theoretical underpinnings of the development of the Evaluation Use and Evaluation Involvement Scales, the definitions of the terms use, influence, and involvement should be clearly differentiated, as these were the constructs of greatest concern to the developers of the sets of scales. The team was originally curious about the relationship between patterns of evaluation use and influence, and evaluation involvement (Lawrenz, King, & Greenseid, 2005). The team was interested in both the relationship between involvement in the evaluation and the use and influence of the evaluation on the evaluation participants, and the factors that appeared to be related to the use

and influence of the evaluation on the projects (King, et al, 2007). The definitions used during the creation of the two sets of scales are presented in Table 1.

Table 1. *Definitions of Use, Influence, and Involvement*
(Lawrenz, King, & Greenseid, 2005)

Term	*Definition*
Evaluation Use	The purposeful application of evaluation processes, findings, or knowledge to produce an effect
Influence on evaluation	The capacity of an individual to produce effects on an evaluation by direct or indirect means
Influence of evaluation	The capacity or power of evaluation to produce effects on others by intangible or indirect means
Evaluation involvement	Active engagement in at least one phase of an evaluation (planning, implementing, applying)

The following sections elaborate on these four definitions and explain the theories behind the creation of the sets of scales developed to measure both use and involvement in program evaluations.

Evaluation Use

In their book, Foundations of Program Evaluation, Shadish, Cook, and Leviton ask, “What good is a fine evaluation of a program that solves a serious problem if the results are not stored and used to ameliorate the problem?” (1991, p. 21). The authors outline the four stages of evaluative problem solving: “(a) identifying a problem, (b) generating and implementing alternatives to reduce its symptoms, (c) evaluating these alternatives, and then (d) adopting those that results suggest will reduce the problem satisfactorily” (p. 20-21). While all four

stages are critical to the evaluation process, evaluation use is the conclusive step that allows those involved in the evaluation process to learn from the evaluation and implement changes based on their findings.

Three broad categories of use have been identified in the evaluation literature: instrumental use, which involves direct decision making about a program based on the evaluation results; persuasive or symbolic use, where a decision maker attempts to persuade others to take a position he or she already espouses; and conceptual use, or enlightenment, where the results of an evaluation are used to affect or change how people feel about an issue or to educate decision makers (Shadish, Cook, and Leviton, 1991; Weiss, 1998). Not all three types of use may be the goal in every situation, but evaluators and others involved in the evaluation process must be conscious of the type(s) of use that are the desired outcome of their particular evaluation.

Factors Affecting Use

Important to the theoretical frameworks underpinning both the Evaluation Use and Evaluation Involvement Scales, are the factors found by previous researchers to affect the use of an evaluation. Alkin, Daillak, and White (1979) presented a framework of evaluation utilization, their term for use. This framework included a number of factors affecting use. Most relevant to the present study are evaluator credibility, information content and reporting, and political factors both inside and outside the organization. Because some of the techniques used in an evaluation are beyond the expertise of the evaluation user,

he or she must trust the evaluation professional to make decisions related to the evaluation, and this trust gives the evaluator credibility. The structure within and around an organization undergoing an evaluation must also affect the use of the evaluation findings as political factors will be in place at both levels that will encourage or discourage the buy-in of various stakeholders. Those with greater buy-in will be more likely to use the results of the evaluation, but, unless those with higher standing within the organizational structure also choose to use the results, use may be limited. Finally, use is also affected by the way the evaluation results are presented to the intended users of the evaluation. Evaluators have much data, often both qualitative and quantitative, to share with the program stakeholders. The ways in which certain data are highlighted over other data can have an impact on the use of the evaluation findings.

Evaluators concerned with the usefulness of their evaluation process and findings must specifically consider use as they plan and implement the evaluation. According to Patton (2008), utilization-focused evaluation, "begins with the premise that evaluations should be judged by their utility and actual use" (p. 37). His suggestions for fostering use of the evaluation build on those presented in the Alkin, Daillak, and White framework and include identifying the primary intended users of the evaluation and the important role played by an identifiable individual or group of people who care about the evaluation and its findings.

When considering use, it is important to distinguish between use that is planned for, or intended by the evaluator, and use that was not specifically intended. This could be the result of participation in the evaluation process or use of evaluation products by unintended users or by intended users in ways that were not intended. Much of the literature on evaluation use is based on intended users of the evaluation results, for instance Patton's (2008) suggestion that evaluators first identify the primary intended users of their work. In her Integrated Theory of Influence (see Figure 1, below), Kirkhart (2000) presents the idea of intention as a dimension. She cautions evaluators to be conscious of the fact that an evaluation may result in use (she uses the term influence) that is intended, unintended, or a combination of the two, and even suggests that unintended use has the potential to be more impactful than use that is intended. Evaluators, then, would be wise to consider both the intended uses of the evaluation they are planning as well as potential unintended uses. And, as Patton suggests in his discussion of primary intended users, these uses may occur with both intended and unintended users of the evaluation results. In the case of the Evaluation Use Grant research, the National Science Foundation (NSF) was the primary intended user of the evaluations studied by the research team, but whether it was intended that the NSF grantees also use the evaluation results is not as clear, as some of the evaluations were designed to disseminate their findings to the grantees while others were not (Lawrenz & King, 2009). Due to the funding system creating a situation where the evaluations were designed

to mostly meet the needs of the NSF as the primary intended user, the grantees could best be described as a mix of intended and unintended users who intentionally or unintentionally used the evaluation results.

The frameworks presented below build on the previously described foundation of the importance of use in evaluation. The frameworks are presented in chronological order of publication to highlight the historical timeline of creation of the theoretical frameworks used in the development of the Evaluation Use Scales.

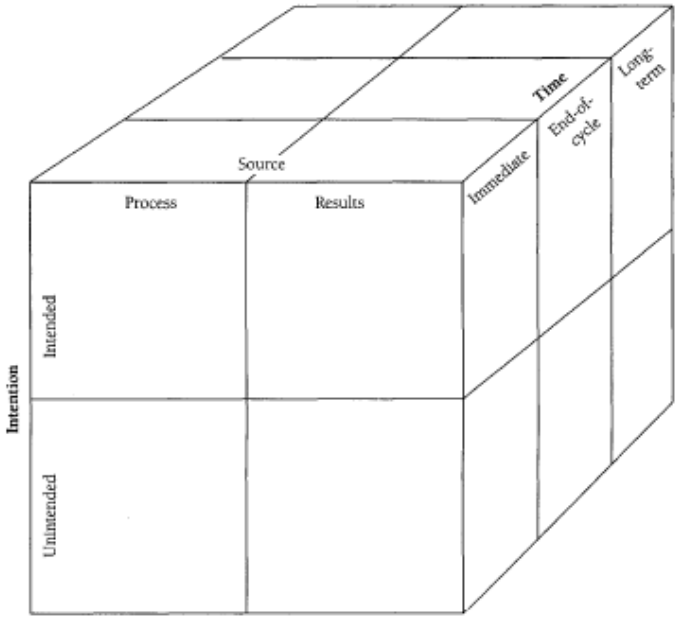
Frameworks for Assessing Evaluation Use and Influence

Shadish, Cook, and Leviton (1991) suggested that evaluators need a theory to tell "how, when, where, and why they can produce useful results" (p. 52). These theories of use so important to evaluators, the authors suggest, have three elements: a description of possible types of use (categorized earlier in this literature review into three main types of evaluation use: instrumental, persuasive, and conceptual), time frames within which use occurs, and specific acts by the evaluator designed to facilitate use of the evaluation findings. These three elements: types of use, timeframe of use, and facilitation of use by the evaluator, are presented by the authors as a mere beginning for the consideration of use of evaluations. This beginning was later elaborated on by future researchers in the field of evaluation.

In 2000, **Kirkhart** proposed that evaluators move beyond use to consider the influence of their evaluations. Drawing on the frameworks of other

researchers in the field, such as that presented by Shadish, Cook, and Leviton, her conceptualization was presented to both map the influence surrounding evaluations and to improve the validity of studies of evaluation influence. Kirkhart's theory, presented visually below, includes a three-dimensional approach to considering evaluation influence. These three dimensions include the source of influence (process or results), the time of influence (immediate, end-of-cycle, or long-term), and intention of influence (intended or unintended).

Figure 1 Integrated Theory of Influence
(Kirkhart, 2000)



Kirkhart argues for the term influence over use because she views it as the broader of the two terms. She defines influence as "the capacity or power of persons or things to produce effects on others by intangible or indirect means" (2000, p. 7) and suggests that this broader term allows evaluators to examine evaluation effects that are "multidirectional, incremental, unintentional, and

noninstrumental, alongside those that are unidirectional, episodic, intended, and instrumental (which are well represented by the term *use*" (p. 7). Within this influence framework, then, Kirkhart suggests that evaluators consider the three dimensions presented in Figure 1. While the graphic depicts each concept as separate from the others, Kirkhart suggests each should be considered more of a continuum. First, evaluators must consider the source of the influence, either the process or the results. Use of results has been the traditional view of evaluation use, as suggested previously by the three categories of use: instrumental, persuasive, and conceptual. In suggesting that evaluators must also consider the use of an evaluation's process, Kirkhart drew from both Patton (2008) and his ideas of utilization focused evaluation, and Greene (1988), who suggested that process-based influence can occur in the three dimensions of cognitive, affective, and political.

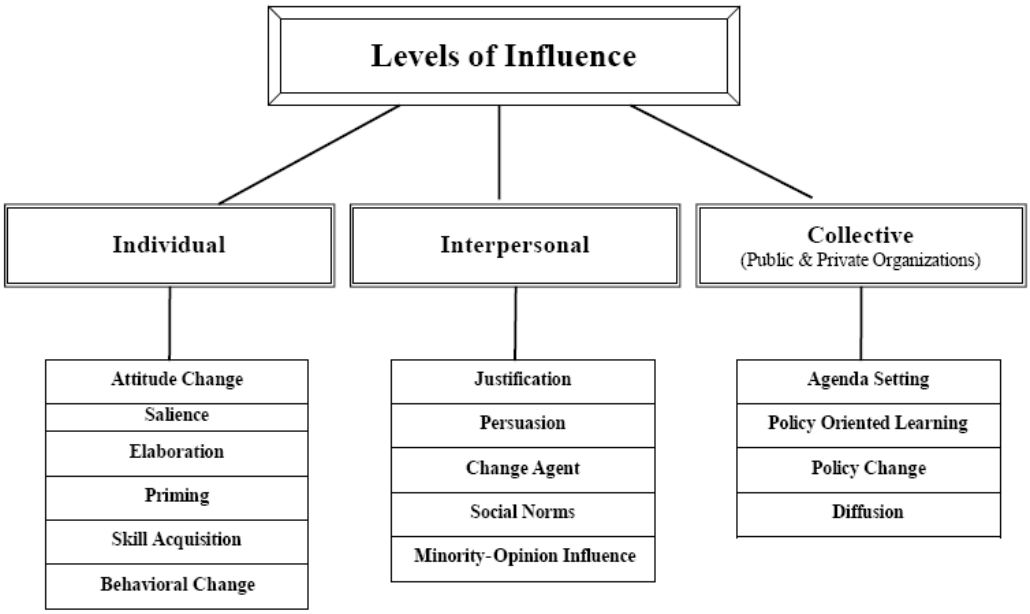
Second, evaluators must consider the intention of the evaluation, which Kirkhart defines as the "extent to which evaluation influence is purposefully directed, consciously recognized, and planfully anticipated" (2000, p. 11). Here a distinction is made between intended and unintended influence and evaluators are encouraged to consider the type of influence they are anticipating, the persons who may be influenced, and the processes, findings, and people who might exert this influence. Finally, evaluators are asked to consider the time dimension of influence, which includes immediate influence that happens concurrently with the evaluation process, as well as end-of-cycle influence, which

occurs at the conclusion of the evaluation process, and long-term influence, which happens long after the evaluation itself has been completed and results have been shared. As the cube-shape implies, these three dimensions should be considered to be interacting with one another so that evaluation influence may be occurring in a situation that involves any of a large number of combinations of factors.

Henry and Mark (2003) built on Kirkhart's redefinition of use as influence and proposed a framework for analyzing the influence of a program evaluation. The first part of their framework, which is visually presented in Figure 2 below, involves classification of change processes and outcomes that evaluations influence at three levels: individual, interpersonal, and collective. The authors define the individual level as, "those cases when evaluation processes or findings directly cause some change in the thoughts or actions of one or more individuals" (p. 297). Henry and Mark suggest that evaluators expect the results of an evaluation to influence an individual's beliefs or opinions, but the process use suggested by Kirkhart implies that an individual may be influenced based on her participation in the evaluation process itself. The second level, interpersonal, refers to "a change brought about in interactions between individuals" (p. 298). This level often involves the use of evaluation findings to persuade others to change their beliefs. The third level, collective, "refers to the direct or indirect influence of evaluation on the decisions and practices of organizations, whether public or private" (p. 298). This level relates back to the political factors both

inside and outside an organization presented by Alkin, Daillak, and White (1979). Within each of these three levels are specific forms of influence that can occur and, as the second part of Henry and Mark's framework suggests, these influence outcomes can be connected to one another through a variety of outcome chains, making it sometimes difficult for an evaluator to trace a specific influence outcome back to a specific aspect of the evaluation process or result.

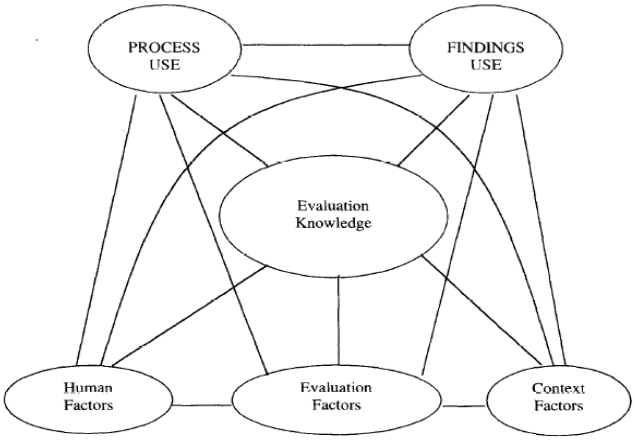
Figure 2 Three Levels of Evaluation Influence
(Henry & Mark, 2003)



Also in 2003, **Alkin and Taut** presented a framework for outlining the distinction between influence and use in evaluation. The authors acknowledge a difference between use and influence, but do not agree with Kirkhart's call to replace the term use with influence. Their framework differentiates between process use and findings use in evaluation. Both types of use, as well as human, evaluation, and context factors contribute to evaluation knowledge and all of

these concepts interact throughout the process of evaluation, as can be seen in Figure 3, below.

Figure 3 Factors Affecting Evaluation Use
(Alkin & Taut, 2003)



The creators of the Evaluation Use Scales used the frameworks of Kirkhart (2000), Henry and Mark (2003), and Alkin and Taut (2003) to inform the creation of the items on their scales, but rather than redefine use as influence, as some of these researchers proposed, the team drew on the work of many evaluation researchers who studied use and influence to create their scales. The Evaluation Use Scales, then, include questions related to use of evaluation findings, influence of evaluations on individuals, and influence of individuals on evaluations (Lawrenz, King, & Greenseid, 2005).

Evaluation Involvement

One main idea behind the creation of the Evaluation Involvement Scales was the notion that involving people in the evaluation process will result in greater ownership in the evaluation which will ultimately lead to more use of the evaluation results. The theoretical framework behind this idea has its beginnings

in the field of participatory evaluation, but goes beyond participation in further defining involvement in evaluation (Lawrenz, King, & Greenseid, 2005). Following is a summary of the relevant research on evaluation involvement.

The Overlap of Utilization-focused Evaluation with Participatory Evaluation

Cousins (2003) defined participatory evaluation as "an approach where persons trained in evaluation methods and logic work in collaboration with those not so trained to implement evaluation activities" (p. 245). Participatory evaluation requires direct involvement in the production of evaluation knowledge by both members of the evaluation community and other stakeholders involved in the specific evaluation being conducted. This evaluation method can be divided into two distinct foci, practical participatory evaluation (P-PE), in which stakeholder participation is expected to enhance the relevance, ownership, and utilization of the evaluation, and transformative participatory evaluation (T-PE), in which empowerment of groups or individuals is the primary function of the evaluation process (Cousins & Whitmore, 1998). Participatory evaluation, particularly the P-PE type, is therefore closely tied to evaluation utilization, or use. This connection is highlighted further in their three-way model of involvement, presented in Figure 4, below.

Recall the previous summary of Patton's (2008) research of Utilization-focused evaluation (UFE) in the section on evaluation use. When defining involvement for the creation of the Evaluation Involvement Scales, the research team first ruled out the UFE idea that involvement happens when primary

intended users are involved in all key evaluation decisions (King, et al, 2009). Then they decided it was not the traditional participatory evaluation definition where participants help to plan and implement the evaluation design. Instead the involvement the team was interested in studying involved key people taking part throughout the evaluation process, which is the point at which UFE and PE overlap. The following frameworks that informed the development of the Evaluation Involvement Scales reflect this overlap between UFE and PE when considering evaluation involvement.

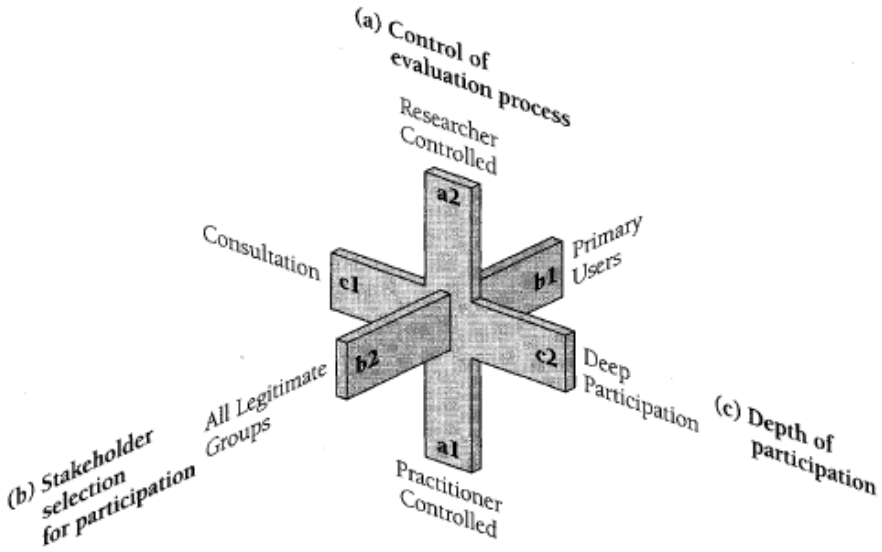
Frameworks for Assessing Evaluation Involvement

As with the research that informed the creation of the Evaluation Use Scales, the frameworks referenced in the creation of the Evaluation Involvement Scales, presented below, build on the previously summarized research on the importance of involvement in evaluation. The involvement frameworks are also presented in chronological order of publication in order to highlight the historical timeline of research used in the development of the Evaluation Involvement Scales.

Cousins and Whitmore (1998) expand their discussion of participatory evaluation to include three distinguishing characteristics that can assist evaluators in assessing participatory evaluations. These characteristics, presented visually below, include control of the evaluation process, stakeholder selection for participation, and depth of participation. The authors acknowledge

that the three dimensions may be independent of one another, but that they most likely interact.

Figure 4 Dimensions of Form in Collaborative Inquiry
(Cousins & Whitmore, 1998)



Burke (1998) suggests a spiral design for the process of participatory evaluation that involves eight key decision points, including: (a) deciding to do the evaluation, (b) assembling the evaluation team, (c) making a plan, (d) collecting the data, (e) synthesizing, analyzing, and verifying the data, (f) developing action plans for the future, and (g) controlling and using outcomes and reports. Though rooted in general participatory evaluation theory, the Evaluation Involvement Scale was based primarily on Burke's process for ensuring meaningful stakeholder participation in the evaluation process. Through this process, Burke explains, "participants develop ownership of the decisions made and develop new skills and confidence to improve their programs" (p. 55).

The team that developed the Evaluation Involvement Scales drew on research in the areas of UFE and PE, as well as Burke's suggested design for the process of participatory evaluation. Involvement was defined based on it being an overlap between UFE and PE. As outlined at the beginning of this paper, involvement was defined to be active engagement in at least one phase of an evaluation, for instance, planning, implementing, or applying (Lawrenz, King, & Greenseid, 2005). Though both the Evaluation Use and Evaluation Involvement Scales were well grounded in evaluation theory, the theories lacked data to offer evidence of their validity as useful theories in this area of evaluation research. The team that designed these sets of scales recognized this need for data collection, and this hole in the research was one of the driving forces behind the creation of the sets of scales.

Measures of Evaluation Use and Involvement

In their review of forty-two studies encompassing twenty years of research on evaluation use and influence, Toal, et. al. (2006) found some explicit testing of theories of use and influence as well as diversity in methodological design, including experimental and quasi-experimental designs, single and multiple case studies, and survey designs. They did not find, however, a measurement instrument designed and validated for use in evaluation settings to quantify the use of evaluation process, findings, and knowledge, the influence of individuals on an evaluation, or the influence of the evaluation on others. In a separate review, Toal (2007) also found a number of studies that measured the

participation of stakeholders in evaluations, but, again, found no universally used measure of the level of involvement of key people throughout the evaluation process.

Consequently, the team of researchers working on the Beyond Evaluation Use Grant developed a set of survey items designed to measure both involvement and use in evaluations. The items designed to measure involvement were grouped into the themes of evaluation planning, evaluation implementation, and communication of evaluation findings. As Toal (2007) describes, subsequent exploratory factor analysis conducted on the initial implementation of the survey instrument with a sample of 369 people involved in evaluations of four different National Science Foundation-funded programs in the field of science, technology, engineering, and math (STEM) education, resulted in the involvement items factoring into two scales, involvement in planning and involvement in implementation. The factors found in the analysis differed in number from those expected by the theory behind the creation of the survey instrument. While the questions were designed to fit into the three themes listed above, the subsequent EFA of the data resulted in the two factors, involvement in planning and involvement in implementation (Toal).

The items designed to measure use of the evaluation were grouped into the themes: influence of the program evaluation on knowledge and understanding; influence of the program evaluation on skills; influence of the program evaluation on personal beliefs; uses of knowledge, products, and

processes from the program evaluation; and ways of using program evaluation findings. These five themes factored into four scales upon exploratory factor analysis of the same survey data used with the evaluation involvement scales. These factors were: influence on knowledge and skills; influence on beliefs; use of knowledge, products, and processes in a future evaluation; and use of evaluation findings. Johnson (2008) completed this exploratory factor analysis of the Evaluation Use Scales.

Confirmatory Factor Analysis and Survey Scale Development

Factor Analysis Defined

The previous discussion of the theoretical underpinnings of the Evaluation Use Scales and the Evaluation Involvement Scales led to the aforementioned attempts by two researchers to gather validity evidence for these sets of scales for use by the evaluation community using exploratory factor analysis. Since the results of these analyses resulted in somewhat unexpected findings, considering the number of factors was not equal to the number expected, the need exists for further exploration of the scales. The next section of this paper will explain the history of the statistical procedure of factor analysis, specifically highlighting confirmatory factor analysis (CFA) and the use of CFA in the field of evaluation as a method for gathering validity evidence for the use of survey instruments as well as a review of the common terms used in describing the process of survey instrument development.

As with the previous discussion of evaluation use and involvement, it will be helpful to first define some of the terms associated with factor analysis. These terms are presented in Table 2 below.

Table 2 *Factor Analysis Definitions*
(Brown, 2006)

Term	*Definition*
Indicator	Observed measures, such as test items or survey or questionnaire items
Factor	An unobservable variable that influences more than one observed measure and that accounts for the correlations among these observed measures
Factor Score	The score that would have been observed for a person if it had been possible to measure the latent factor directly
Common factor model	A postulate that states that each indicator in a set of observed measures is a linear function of one or more common factors and one unique factor
Factor analysis	A statistical technique used to determine the number and nature of latent variables or factors that account for the variation and covariation among a set of observed measures, commonly referred to as indicators
Exploratory factor analysis	A data-driven approach to factor analysis such that, generally, no specifications are made in regard to the number of latent factors or to the pattern of relationships between the common factors and the indicators

Confirmatory factor analysis	*An empirical and conceptual-driven approach to factor analysis that requires the researcher to prespecify all aspects of the factor model including the number of factors, the pattern of indicator-factor loadings, and other aspects*

Historical Background of Factor Analysis

The concept of factor analysis was first introduced by Spearman (1904) and in the more than one hundred years since its first use as a multivariate statistical procedure, it now appears to be ubiquitous in applied research areas including psychology, education, sociology, management, and public health (Brown, 2006). As currently used, factor analysis is based on a postulate described by Thurstone (1935) as "the standard scores of all individuals in an unlimited number of abilities can be expressed, in first approximation, as linear functions of their standard scores in a limited number of abilities" (p. 50). Given this postulate, the process of factor analysis in general involves partitioning of the variance of each indicator into common variance (accounted for by the latent factor) and unique variance, which includes both variance specific to the indicator and random error (Brown). It is possible for factor scores to be computed by simply averaging or summing the values of the items measured. These scores are generally called coarse factor scores and are problematic in that their representation of the latent variable has not been tested. Factor scores calculated through multivariate methods, such as factor analysis, are called refined factor scores and, through the process, become scores that more accurately reflect the latent factor which they are purported to measure (Brown).

Factor analysis can be used for construct validation and data reduction (Brown) and is thought to be the most commonly used procedure in the process of developing latent variable measures (Floyd & Widaman, 1995).

It is important to understand the difference between the two types of factor analysis and how they are most often used. Exploratory factor analysis (EFA) is commonly used in data reduction situations, where researchers with large sets of scores from intercorrelated indicators wish to reduce these scores to factor scores that can then be used as the units in statistical analyses of the data (Brown, 2006). Applied researchers can also use this technique in the early stages of instrument development, such as surveys (Floyd & Widaman, 1995). Since it is data-driven, as defined in Table 2, researchers might explore the usefulness of a survey under development by using EFA to examine whether the single construct being measured by the multiple items account for the intercorrelations of the items, which provides evidence that a single construct is indeed being measured. The researcher can also use EFA to assess whether the items in question are indicators of the construct under study by observing the various loading of the items on the single factor, or construct (Brown). As indicated by its name, EFA is generally used at the exploratory stages of instrument development, before formal validation of the use of the instrument begins.

Confirmatory factor analysis (CFA), on the other hand, is used to test a priori hypotheses (Floyd & Widaman, 1995). CFA requires the researcher to

have prior empirical or conceptual evidence that certain factors do exist and together create a single, construct, or multiple, related constructs. The analysis provides validation evidence for that construct (Brown, 2006). As explained by Floyd and Widaman, “construct validity is supported if the factor structure of the scale is consistent with the constructs the instrument purports to measure” (p. 287). EFA is often conducted prior to CFA in the development of an instrument so that the factors found during the exploration stage might be confirmed with a new sample of data in the confirmatory stage (Brown; Jöreskog, 1969). Along with confirmation of constructs, CFA can be used to compare the fit of competing hypothesized models, and the process can also suggest modifications or alterations to factor structures that can lead to better model fit (Floyd & Widaman). CFA is a particularly important tool in the social sciences because of the prodigious use of questionnaires as assessment tools in the field (DiStefano, 2002).

The Development of Survey Items and Instruments

An attempt to define the terms commonly used in describing the creation of survey instruments led to the definitional confusion mentioned previously in the introduction to this review of the literature. Table 3 presents a sampling of the variety of definitions used by authors to describe some of the common terms in survey development. Table 4 summarizes the definitions this researcher has chosen to use in order to provide consistency in this section of the review of the literature.

Table 3 *Definitions Related to Survey Scale Development*

Term	*Definition*
Scale (Nardi, 2002, p. 57)	A set of items that are ordered in some sequence and that have been designed to measure a unidimensional or multidimensional concept.
Scale (Carifio & Perla, 2007, p. 112)	A purposely constructed (according to an a priori blue print and plan) interrelated set of items which have defined and targeted logical and empirical properties.
Likert Scale (Carifio & Perla, 2007, p. 113)	A series of verbal statements that expressed a range of positive expressions, views, sentiments, claims or opinions about the attitude object (underlying construct) that ranged from mildly positive to strongly positive and then the same relative to a range of negative statements.
Likert Scale (Jamieson, 2004, p. 1217)	Commonly used to measure attitude, providing a range of responses to a given question or statement, typically there are 5 categories of response from (for example) 1=strongly disagree to 5=strongly agree.
Intensity Scale (Nardi, 2002)	Ordinal survey items that measure level of intensity, such as a 5-point range from strongly disagree to strongly agree.
Summative Attitude Scale (Thorndike, 2005)	Responses to statements using numerical indication of the strength of the subject's feeling toward the object or position described in the statement. The responses are then summed to estimate the strength of the attitude.

As can be seen in Table 3, there is little agreement in the literature for the definition of the term scale. Nardi differentiates the term scale from index in stating, "usually, a pattern is sought from the responses to a set of items, rather

than a simple summation of the individual item scores, as with indexes" (2002, p. 57). He goes on to say that most researchers are constructing indexes, "a set of items that measure some underlying concept" (p. 56), when they indicate they are constructing scales. Things are further confused when terms such as Likert scale, rating scale, or attitude scale are used to refer to ordinal measurement of a single item on levels such as agreement/disagreement. As Carifio and Perla (2007) point out, much of the confusion seems to be due to the use of the term scale grammatically as a subject, predicate, adjective, and process. The authors state that scale can be both a collection of items and the manner in which the items were measured, such as the use of the term Likert scale by Jamieson presented in Table 3. The definitional issues seem to be at the heart of the controversy surrounding the argument for and against treating scales as ordinal or interval variables in statistical analyses. The next section of this review of the literature delves more deeply into this controversy and what it means for projects using ordinal variables, such as the confirmatory factor analysis of the Evaluation Involvement and Evaluation Use Scales. The definitions that coincide with the terms that will be used in the next section appear below in Table 4.

Table 4 *Definitions of Terms to be Used in Reference to the Evaluation Involvement and Evaluation Use Scales*

Term	*Definition*
Survey	Collecting information for the purpose of producing statistics by asking questions to a fraction of the population – a sample – rather than from every member of the population (Fowler, 1993)
Scale	A set of items that measure some underlying shared concept; items share a pattern and are not simply summed (adapted from Nardi, 2006 definitions of both index and scale)
Intensity Response Format	The measurement method used with ordinal survey items that measure the intensity of something, such as levels of agreement/disagreement or levels of involvement in or use of an evaluation (adapted from Nardi, 2006 definition of intensity scale)
Survey Item	A single question on a survey. In the case of the Evaluation Use and Evaluation Involvement Scales, The survey includes 6 scales. There are 43 items that make up these 6 scales that are all measured with an intensity response format.

The definition presented in Table 4 for survey is a common example of those found in the literature and does not appear to be associated with controversy. Scale, on the other hand, as presented above and further discussed later in this paper, is rife with controversy, much of it stemming from the confusion related to “Likert scales”. The use of scale by some to mean a single item measured using an intensity response format and others to mean a set of items that may or may not be measured in this way, has caused much of the disagreement regarding data analytic methods that are appropriately used

with this type of data. The term index might have been chosen, in order to remove scale completely from discussion, but scale is a commonly used term and scaling as a method is common practice in measurement, so it was decided to use the term, but only in the context of the combination of a group of items that, in the case of the Evaluation Use and Evaluation Involvement Scales, contain items measured with intensity response formats. The confusion over the use of the term Likert scale (Jamieson, 2004; Carifio & Perla, 2007) and the fact that the term Likert scale or Likert-type response format is often generalized too far, as Rensis Likert developed items measuring attitude in a very specific way (Likert, 1932), led to the selection of the term intensity response format to describe the types of items created for the Evaluation Use and Evaluation Involvement Scales. This term is more generalizeable and can be related to many types of items measured with categories of intensity, of which Likert-type response formats is one example.

Issues Related to the Use of Ordinal Data

The Disagreement Defined

The disagreement mentioned in the previous section regarding the definitions of terms related to ordinal variables in surveys leads to potential misunderstanding by researchers about the use of various statistical techniques on data obtained from scale administration as well as information provided in a single scale item. The disagreement can potentially be summarized as these four issues:

1. Disagreement on the use of the term scale, as presented in Table 3 (Carifio & Perla, 2007).
2. Disagreement on rules for determining whether a particular item or scale should be considered an ordinal variable (Knapp, 1990).
3. Disagreement on which types of statistical analyses are appropriately completed on ordinal variables (Baker, Hardyck, & Petrinovich, 1966; Labovitz, 1967; O'Brien, 1979; Marcus-Roberts & Roberts, 1987; Knapp; Jamieson, 2004; Carifio & Perla).
4. Disagreement about appropriate data analysis of items versus scales (Jamieson; Carifio & Perla)

Since the first disagreement has already been summarized, the following section will provide a summary of each of the remaining three disagreements in turn.

Rules for determining whether a particular item or scale should be considered an ordinal variable

In 1946, summarizing the work of a committee of the British Association for the Advancement of Science, Stevens outlined the types of response formats used in measurement, listed in Table 5 as types of variables, and referred to by Stevens as scales, which provides further evidence of the confusion of definitions in the literature. Knapp (1990) believes that the literature lacks an agreed upon set of rules for determining ordinality of scales. It should be noted that Knapp does not specifically define his use of the term scale, but based on the description and examples he gives, the reader can safely assume he means an

item using an ordinal response format when he refers to an ordinal scale. His argument against the existence of an agreed upon set of rules is based on the fact that there would be disagreement by researchers about the ordinality of an item using an intensity response format based on the words used to define each level of order, such as the categories: never, seldom, frequently, and always.

Table 5 *Types of Variables*
(Nardi, 2006)

Term	*Definition*
Discrete variable	A variable that has values that do not contain additional information between those values
Continuous variable	A variable that has values with information between those values
Nominal variable	Discrete measures whose values represent named categories of classification
Ordinal variable	Discrete measures whose values are in sequence, that is they increase or decrease in a particular order
Interval variable	Measures whose numbers are in order with equal size intervals, but have no absolute or fixed zero starting point (usually continuous, though discrete are also possible)
Ratio variable	An interval variable with an absolute zero starting point

A second issue related to the consideration of ordinal variables presented by Knapp is the item's normality. He cites authors who both present the case that a normally distributed variable should be considered interval and those who argue that normality does not automatically equal an interval measure and that an ordinal variable can potentially be normally distributed, which does not change the fact that it is still an ordinal variable and should be treated as such. Whether

there remains disagreement in the empirical literature, there does seem to be evidence of disagreement in the applied literature as to whether an item, at least one measured using an intensity response format, should be considered ordinal or interval, as Nardi (2004) states that it is common practice for researchers to treat items using an intensity response format as interval variables when they believe the intensity level varies in equal intervals along the measure. Whether each researcher's argument is convincing enough to those reading the results, one suspects, is considered one case at a time and perhaps depends on the statistical analyses conducted on the data generated from these ordinal measures. There are also researchers presenting quantitative methods to rescale ordinal data into interval data using item response theory (Harwell & Gatti, 2001). This rescaling does not guarantee normality, since the variable being rescaled may not be normally distributed to begin with, but, if it is, the rescaling may allow the assumption of normality to be met. Leaving the data in ordinal form guarantees that it cannot be considered to be normally distributed (Harwell & Gatti).

Types of statistical analyses appropriate to complete on ordinal variables

There seems to be variance in application of statistical analyses to ordinal variables which possibly stems from continuing disagreement regarding what types of statistical analyses are appropriately conducted on these types of variables. Knapp (1990) describes two categories of opinions on the topic, labeling them pro-Stevens (conservative) and anti-Stevens (liberal) after Stanley

Smith Stevens. In his 1946 summary of the aforementioned committee discussion, Stevens suggested that the only permissible statistics to be used when analyzing data generated by ordinal measurement are median and percentiles. He reiterated this position again in a paper on the topic in 1955. O'Brien (1979), a Knapp-labeled pro-Stevens conservative, used simulations to show the danger of trusting the results of correlations calculated with ordinal measures, showing that skewness of the underlying distribution of the latent variable being measured results in questionable correlation values when the data are treated as interval. Marcus-Roberts and Roberts (1987) further clarified, or muddied, depending on one's outlook, the argument by suggesting that the calculation of means of ordinal variables is appropriate, but that these values may not be meaningful other than to make statements about which number on the intensity response format item is greater than another. Comparisons between respondents or populations may not be meaningful, which impacts the analyses that may be completed on the data. For the conservatives, rescaling would be a necessity for completing statistical analyses of ordinal data that requires the assumption of normality to be met, unless the researcher chooses to use methods specifically designed for use with this type of data, such as "nonparametric procedures, contingency table analysis, regression models for ordinal data, and specialized SEM models" (Harwell & Gatti, 2001, p. 113).

The group labeled anti-Stevens, or liberal, by Knapp suggests that ordinal scales can often be treated the same as interval scales when computing various

statistics, such as Student's t (Baker, Hardyck, & Petrinovich, 1966). Labovitz (1967) found that "arbitrary assignment, which is consistent with rank order, rarely alters the results of statistical analysis to an appreciable degree" (p. 153). Again, the confusion over the use of the term scale is important here, as Baker, Hardyck and Petrinovich used scales as defined in Table 4, that is a set of multiple items that result in one number representing the construct, while Labovitz used single item "scales", which were really items using an intensity response format to measure reactions to therapy on a 4-point rating system. Though the disagreement began more than fifty years ago, there remains confusion in the literature about the appropriate statistical analyses that can be conducted on ordinal scales and ordinal survey items, at least practically speaking. There may be some agreement in the quantitative literature that ordinal data cannot meet the assumption of normal distribution necessary for many statistical procedures, but it appears that applied researchers are treating ordinal data as interval data regardless. In their 2001 paper, Harwell and Gatti surveyed three prominent educational research journals: *American Educational Research Journal, Sociology of Education,* and *Journal of Educational Psychology* to see how often ordinal data were used in research and what statistical procedures were used with these data. The authors found that eighty-five percent of the studies used hierarchical linear modeling or structural equation modeling and that seventy-three percent of the dependent variables in these studies were ordinal data using "Likert scales".

Appropriate data analysis of items versus scales

The final question is an extension of the previous two: if there were agreed upon rules for appropriate statistical analyses for ordinal variables, does it matter whether those variables are measured using a single item with, for instance, an intensity response format, or measured as a scale using a series of such items to generate a value that measures one construct? As presented previously, those in the liberal camp used both types of ordinal variables in their review of appropriate statistical analyses (Baker, Hardyck, & Petrinovich, 1966; Labovitz, 1967). Again, there appears to be some disagreement in the literature as well as in general application.

Structured rules regarding the use of ordinal data measures either by individual items or scales of items is difficult to find in the applied literature. For example, a brief search of the literature in one research area (higher education) using one national survey with items using intensity response formats (the National Survey of Student Engagement), shows differences in the analyses of items versus scales. Kuh and Umbach (2004) completed a multi-level model using scales from the survey as independent variables and a purposefully chosen set of single items from the survey combined to create the dependent variable, thus mixing analysis of items with analysis of validated scales. Mark and Boruff-Jones (2003), on the other hand, chose five individual items from the survey and correlated them with various other indictors to test their hypothesis related to literacy. Deeper searches in the literature in different research areas

using different surveys would likely yield even more varied treatment of items and scales using intensity response formats.

Issues Related to the Confirmatory Factor Analysis of Ordinal Data

CFA Assumptions and Methodology Decisions

As with any statistical analysis technique, CFA comes with its own set of assumptions. These assumptions include:

- Proper specification of the model
- Measurement errors are independent of latent factors
- Measurement errors are uncorrelated with one another (Flora & Curran, 2004).

Additional assumptions enter the picture when a researcher makes a decision regarding the type of method for estimating the parameters in the analysis. The most common method is maximum likelihood (ML) parameter estimation (Flora & Curran). Other decisions related to the appropriate use of CFA in the validation of surveys include choice of correlation measures, minimum sample size of completed surveys, the size of the a priori model being tested, as well as the number of indicators per factor in this model, and the fit indices used to assess the fit of the model. Recent developments in the study of these many decisions will be summarized later in this paper.

Ordinal Variables

An issue with the use of CFA in the validation of survey instruments in the social sciences is the use of ordinal variables at the item level. As DiStefano

(2002) states, “The Likert response technique is widely used with questionnaires because it is fairly easy to develop, has been shown to be highly reliable, and is adaptable to many situations” (p. 327). DiStefano uses the term “Likert response technique” to indicate an item that includes five choices from which the survey respondent selects her answer. However, as outlined previously in this paper, there is disagreement in the literature about the use of Likert’s name to describe these types of items, particularly when the term “Likert scale” is used (Jamieson, 2004; Carifio & Perla, 2007). Though the disagreement lives on in the literature and in application, the important point to be made related to their use in questionnaires being validated by CFA is that intensity response format items, those commonly called Likert-type items, or sometimes Likert scales in the literature, are items that contain a number of ordered choices (e.g. five categories ranging from 1-strongly disagree to 5-strongly agree) from which the survey respondent selects his answer while completing the survey. As mentioned previously, these types of items are used frequently in social science research, so it is important to consider how their use might impact the process of CFA.

Specific to the Evaluation Use Scales and Evaluation Involvement Scales being potentially validated with CFA, the argument that the intensity response format utilized for each item of the survey scales vary at equal intervals would be a difficult one to make. The items required survey respondents to answer either, “No”, “Yes, a little”, “Yes, some”, or “Yes, extensively” to various questions

related to use and involvement in the evaluation process. While it may have been the intention of the scale creators that these four points of the measure were equidistant, the nature of latent variables and the potential variety of interpretations of the points of the intensity response format leaves a question as to whether the items can individually be treated as interval variables. With their existence as interval variables in doubt, it would arguably be best to assume them to be ordinal variables and find a way to handle them as they are in the CFA process.

The previously outlined pro-Stevens, anti-Stevens disagreement over the appropriate statistical analysis of ordinal variables is also important when considering CFA of these types of variables. The importance to CFA is highlighted when considering that Knapp (1990) summarized the controversy in two parts: first, that there exists no set of rules for deciding whether a variable is ordinal, particularly when terms are used to indicate each point, such as "occasionally" or "sometimes". Second, there is disagreement about whether the shape of a variable's distribution impacts its ordinality. Some argue that normality automatically constitutes intervality, but others disagree (Knapp). In the applied literature, Nardi (2004) states that it is common practice for researchers to treat items using an intensity response format as interval variables when they believe the intensity level varies in equal intervals along the measure. While these arguments may seem like so much semantics, a survey item's

format and its accompanying assumptions of normality are important factors in the appropriate methodology choices made in the completion of CFA.

Issues associated with completion of CFA with ordinal variables

Based on the above description of intensity response format items and ordinal variables, it is clear that problems arise when completing CFA with these items. As DiStefano states, when analyzing intensity response format items,

> The observable level representing the latent construct is crude in nature. The crudeness arises from cutting the continuous scale of the construct into a set number of ordered categories and error is introduced into analyses from the imperfection of the scaling technique (2002, p. 328).

Using these crudely ordered variables leads to potential problems in CFA due to the common application of the maximum likelihood method for estimating parameters in the analysis process (Flora & Curran, 2004). In addition to the assumptions for CFA in general, listed previously in this paper, maximum likelihood estimation includes assumptions of its own. The two important assumptions related to the use of ordinal variables are that the computation of the sample covariance matrix is completed with data that is 1) continuous, and 2) multivariate normal in distribution (Flora & Curran).

The first assumption highlights the importance of the disagreement in the field regarding the ordinality or intervality of intensity response format items summarized earlier in this paper. This assumption is violated when analyzing these types of ordinal items because, though designed to measure a continuous

latent variable, item-level responses to these intensity response format items are actually discrete because only a small number of categories are listed (e.g. 1-strongly disagree to 5-strongly agree) (Flora & Curran). The second assumption is also violated because discrete variables cannot truly be normally distributed. Previous studies have shown that ordinal data, particularly those items that have five or fewer categories, produce biased results when used in maximum likelihood CFA (Babakus, Ferguson, & Jöreskog, 1987; Muthén & Kaplan, 1985; Green, Akey, Fleming, Hershberger, & Marquis, 1997; Hutchinson & Olmos, 1998). While researchers have successfully outlined the problems with CFA of ordinal data, they have also presented some solutions, which are summarized below.

Solutions to the issues related to the CFA of ordinal data

The solution for the problems related to the use of ordinal variables in CFA is not as simple as selecting a different method of parameter estimation other than maximum likelihood, which has been shown to produce biased results in this situation. The myriad issues listed previously in this paper all impact the choices made when planning a CFA and they interact with one another to compound issues as well. Potential solutions are presented below in sections in order to highlight the research that has been done related to the impacts of ordinal variables on the different aspects of the analysis process in CFA.

Methods for estimating parameters

The first set of solutions for solving the issues related to ordinal variables and CFA involves the selection of an appropriate method for estimating parameters. The impact of lack of continuous, multivariate normal data resulting from ordinal variables on maximum likelihood estimation has been mentioned previously and will be elaborated on here. A number of studies have examined the efficacy of different estimation methods.

Muthén and Kaplan (1985) designed a study to compare four methodologies for estimating parameters in the CFA of five-category intensity response format items: maximum likelihood (ML), generalized least squares (GLS), asymptotically distribution-free generalized least squares (ADF), and categorical variable methodology (CVM). The authors chose to compare the treatment of the data as interval scale normal with interval scale non-normal and specifically considered issues of discreteness, skewness, and kurtosis in their Monte Carlo simulation study. Results indicated that, while ADF and CVM seem to perform better than ML and GLS when data show strong skew and/or kurtosis, they are not useful in studies with more than moderate numbers of factors (25-30). The authors recommended both ML and GLS as methods when skewness and kurtosis are less than 0 or correlations are $\leq .2$. Further, when sample size is less than 400, ML is recommended over GLS. Their follow-up study (1992) further assessed the use of GLS with what they felt was a more realistic sample size, 500 as compared to a less obtainable sample size of 1000, finding that the

chi square estimates and standard errors of GLS are not as robust to non-normality with a sample size of 500.

DiStefano (2002) compared ML to the weighted least squares (WLS) approach to estimating parameters. In addition, she specified the measures of correlation used with each method: Pearson product moment (PPM) in the case of ML, and polychoric correlation (PC) in the case of WLS. To further clarify the methods debate, she also attempted the Satorra-Bentler rescaling correction with the ML method to further account for the ordinality of the data. Results showed that the ML-PPM method resulted in negatively biased parameter estimates, though the Satorra-Bentler rescaling procedure did satisfactorily alleviate the bias. The WLS-PC method was robust to non-normality of data, but did show biased estimates when sample sizes were small in relation to model size. This study confirmed the findings of Muthén and Kaplan (1985) by recommending least squares parameter estimation, though it specified the use of WLS and also offered ML with rescaling as an option when small sample size prohibits the use of WLS.

Flora and Curran (2004) went beyond DiStefano's study by comparing WLS with robust WLS as methods for estimating CFA parameters with ordinal variables, such as intensity response format items, using polychoric correlation with both WLS methods. While both methods showed little bias in parameter estimation given moderate levels of non-normality, robust WLS appeared to be unbiased at most sample sizes ≥ 200. WLS had more issues with smaller

samples. The authors did note, however, that their study did not provide a thorough assessment of the effects of more extreme skewness and kurtosis.

Finally, Jöreskog and Moustaki (2001) introduced the assessment of three other approaches to parameter estimation in CFA: the underlying bivariate normal approach (UBN), the normal ogive approach (NOR), and the proportional odds model approach (POM). Since no structural equation modeling software includes these methods, the authors wrote their own programs to test their efficacy in parameter estimation in CFA with ordinal data. While the results were somewhat promising, the authors concluded that none of the approaches were practical for applied use and therefore, like DiStefano (2002) and Flora and Curran (2004), recommended WLS with polychoric correlation for those attempting CFA with ordinal variables.

It would appear, based on these studies of the various parameter estimation methods, that WLS, and specifically robust WLS, is the method of choice for unbiased parameter estimation of moderately non-normal, ordinal data in the process of CFA, at least when sample sizes are 200 or greater.

Measures of correlation

Due to the design of a number of the previously summarized studies, it is difficult to separate the presented solutions for selecting appropriate measures of correlation from the selection of parameter estimation methods. To summarize, both DiStefano (2002) and Flora and Curran (2004) recommended the use of polychoric correlation with the WLS parameter estimation method. Prior to this

apparent consensus in the literature for the use of WLS as a parameter estimation method, however, Babakus, Ferguson, and Jöreskog (1987) completed a study assessing the sensitivity of four measures of correlation to ordered categorical data. Using the ML estimation method, they tested the product-moment, polychoric correlation (PC), Spearman's rho, and Kendall's tau-*b*. The authors found that PC performed the best on the basis of factor loading bias and squared error, standard errors, and accuracy of parameter estimates. However, PC also produced the poorest fit statistics of the four correlations in the study. The authors recommended PC as the best correlation measure with the caveat that it does not properly measure fit. They further suggested that PC be studied with least squares estimation, just as it was by DiStefano and Flora and Curran.

Fit indices

An important aspect of CFA is the resulting measures of model fit produced by the analysis. It is recommended that at least one index from each of three fit classes: absolute fit, comparative fit, and parsimony, be considered when determining the fit of the CFA model in question (Brown, 2006). As with parameter estimation methods and correlation measures, researchers have a choice of which fit indices in each category to use when completing CFA and, according to the literature, some are more or less impacted by the ordinality of the data being analyzed.

Hutchinson and Olmos (1998) used a Monte Carlo simulation to test the effects of sample size, model size, estimation procedure, and level of normality on ordinal data in CFA. Based on previous research by Babakus, Ferguson, and Jöreskog (1987), Sharma, Durvasula, and Dillon (1989), and Wang, Fan, and Wilson (1996), the authors felt that there was evidence showing that the χ^2 test of fit is biased when used with non-normal data. They included the χ^2 test in order to compare results with these previous studies, but also tested the comparative fit index (CFI), critical N (CN), incremental fit index (IFI), measure of centrality (MOC), nonnormed fit index (NNFI), relative fit index (RFI), and root mean square error of approximation (RMSEA). Results agreed with previous studies, showing that the χ^2 test was negatively impacted by non-normal data, though use of WLS with PC did somewhat alleviate this negative impact. The authors suggested the use of RMSEA when used in tandem with WLS with PC as it performed well under varied sample and model sizes. CN was found to be the only index of the eight tested that was not impacted by non-normality, though it did have other flaws. Therefore the authors suggested that it might be used in tandem with other fit indices when data are non-normal. Finally, NNFI was recommended for use based on the fact that it appeared to be free from effects of sample size, but was also more sensitive to model misspecification than were some other indices.

Related issues: Sample size, model size, indicators per factor, and categories per item

Though not specifically related to ordinal data, some other issues to consider when completing CFA became apparent in the summarization of the studies reviewed for this paper. Many of the studies cited in this review of the literature tested multiple CFA issues at once (Babakus, Ferguson, and Jöreskog, 1987; Hutchinson & Olmos, 1998; DiStefano, 2002; Flora & Curran, 2004), highlighting the possible interaction effects of the various choices made when conducting CFA. Choice of parameter estimation methods, correlation measures, and fit indices are not only related to the non-normality of data from intensity response format items, but also to the sample size of the dataset, the size of the model being fit, the number of indicators per factor, and even the number of categories in the items being factored. Muthén and Kaplan (1992) highlighted the importance of sample size in the follow-up to their 1985 study on methods of parameter estimation in CFA, finding that sample sizes of 500 impacted robustness to non-normality. Olmos and Hutchinson recommended using WLS with PC and the RMSEA fit index, but only when the sample size is large enough to estimate the weight matrix necessary for WLS. Babakus, Ferguson, and Jöreskog found an effect of sample size on fit indices. Dolan (1994) highlighted the potential interaction between sample size and the number of categories per item showing that, for proper estimation of χ^2 and Pearson Product Moment Correlation, five response categories are a minimum for obtaining robust results, given small (200 to 400) sample sizes. Breaking somewhat from the findings of other researchers, he also cautioned against the

use of polychoric correlations to account for the non-normality of ordinal data, stating that it is likely to lead to misjudgments.

DiStefano emphasized the sample size/parameter estimation method interaction as well as the possible effects associated with the fitting of a large model. Potthast (1993) also expressed caution related to the interaction between sample size and model size, noting that CFA with ordinal data on a sample size <1000 is likely to be suspect, particularly when the model size is large (the largest model in her simulation study being 16 variables and 4 factors). With high skewness or kurtosis, the probability of negative bias of standard errors and goodness of fit indices are greater with smaller sample sizes and larger model sizes. In their study, Flora and Curran highlighted the interactions among parameter estimation method, data non-normality, fit indices, and number of item categories. Based on the findings of these studies, and others cited by the authors of these studies, it is apparent that a set of recommendations for choice of parameter estimation method, correlation measures, and fit indices cannot be made without study-specific consideration of other factors, including sample size, model size, number of indicators per factor, and number of categories per item when completing CFA on ordinal data.

In their meta-analysis of thirty-four CFA studies, Hooglund and Boomsma (1998) found both contradictory results regarding the methodological issues related to CFA presented in this literature review as well as a lack of realism in models studied by empirical CFA researchers. Hutchinson and Olmos (1998)

also lamented the lack of CFA research on ordered categorical data, which are so commonly used in social science questionnaires. It seems, however, based on the more recent CFA studies, that researchers are attempting to make their simulations more representative of data encountered in the field (DiStefano, 2002; Flora & Curran, 2004).

Table 6 *Summary of Solutions to Issues with CFA of Ordinal Variables*

Issue	*Solution*	*Citation*
1. Parameter Estimation Methods	ADF and CVM better with skewness or kurtosis; ML or GLS better when skewness or kurtosis < |1.0| and with >25-30 factors; ML recommended when N<400	Muthén & Kaplan, 1985
	Recommend WLS with polychoric correlation	Jöreskog & Moustaki, 2001
	Recommends WLS unless N is small in relation to model size, then use ML with Sartorra-Bentler rescaling technique	DiStefano, 2002
	Recommend robust WLS with N≥200, but this study did not account for skewness or kurtosis of data	Flora & Curran, 2004
2. Measures of Correlation	Recommend polychoric correlation with WLS parameter estimation	DiStefano, 2002 Flora & Curran, 2004 Babakus, Ferguson, & Jöreskog, 1987
	Caution that the method does not properly measure fit	Babakus, Ferguson, & Jöreskog, 1987
3. Measures of Fit Indices	Recommend: RMSEA (with WLS-PC); Critical N (in tandem with other measures to account for its flaws);	Hutchinson & Olmos, 1998 (based on: Babakus, Ferguson, &

	and Non-normed Fit Index	Jöreskog, 1987; Sharma, Durvasula, & Dillon, 1989; Wang, Fan, & Wilson, 1996)
4. Interaction of Issues	**Sample Size:** χ^2 and standard error estimates of GLS not as robust to non-normality with realistic (N=500) sample size	Muthén & Kaplan, 1992
	Sample Size and Number of Categories per Item: To properly estimate χ^2 and PPM correlation, recommend minimum of 5 response categories per item; cautions that use of polychoric correlation can lead to misjudgments	Dolan, 1994
	Sample Size and Number of Indicators per Factor: Cautions against use of CFA of ordinal data with N<1000; High skewness or kurtosis increases probability of negative bias in standard errors and goodness of fit indices with large model sizes (16 variables and 4 factors)	Potthast, 1993

Conclusion

The use of the term scale to indicate different aspects of surveys has led to confusion for researchers. The literature continues to vary on the treatment of ordinal variables as interval variables both by definition and application. Further confusing the issue is the variety of statistical analyses conducted on both items and scales using intensity response formats.

Beyond the confusion related to the definition of scale and the treatment of ordinal variables in statistical analysis, there does appear to be a strong argument for the use of factor analysis in survey development. Brown (2006) suggests that factor scores, scores that a person would demonstrate were it possible to directly measure the latent factor under consideration, can be calculated in different ways. As mentioned previously, coarse factor scores can be computed simply by averaging or summing the items thought to be related to the factor being considered. Refined factor scores, which can be calculated using multivariate methods, such as Thurstone's least squares regression approach, or factor analysis, are potentially less biased than their coarse counterparts. A problem with refined factor scores, though, is that "there is an infinite number of sets of factor scores that could be computed from any given factor analysis that would be equally consistent with the same factor loadings" (Brown, p. 37). Exploratory and confirmatory factor analysis can both examine and account for this so called indeterminacy, but the researcher has to purposefully look for this problem and accommodate it in the analysis. It should also be noted that EFA and CFA can produce discrepant results, and interpreting these potential differences is part of the process of validating survey scales (Floyd & Widaman, 1995).

In regard to CFA with ordinal variables, while a concise set of recommendations for choice of parameter estimation methods, correlation measures, and fit indices cannot be provided for the reasons summarized

previously in this paper, a concise summary of the relevant findings is in order. With parameter estimation methods, it is apparent that maximum likelihood estimation is more prone to bias with non-normal data, and ordinal data is likely to be non-normal, sometimes showing moderate to extreme skewness and/or kurtosis. The weighted least squares method is more robust to non-normality, particularly robust WLS, but has been found to be more sensitive to sample size. Researchers are advised to consider sample size when selecting the appropriate parameter estimation method for their application of CFA and should consider the importance of sample size considerations when planning their study and especially prior to data collection. If a sufficient sample size is not realistic, then perhaps CFA should not be attempted on the data, though what constitutes a sufficient sample size is still under debate in the literature (Hutchinson and Olmos, 1998; DiStefano, 2002; Flora & Curran, 2004).

There appears to be more agreement regarding the appropriate choice of correlation measures when conducting CFA. When a researcher possesses a large enough sample size to conduct a CFA with WLS (or robust WLS) estimation, then polychoric correlation is recommended. Polychoric correlation is also recommended when using ML, and should the researcher need to use this method due to small sample size, she should use the Satorra-Bentler scaling technique to account for the non-normality of the data.

Fit indices that have been recommended based on empirical evidence include the root mean square error of approximation, when used in tandem with

weighted least squares parameter estimation and polychoric correlation, nonnormed fit index when sample size is relatively small, and critical N due to its freedom from negative impact due to non-normality. As mentioned previously, however, it is critical for researchers using CFA to consider sample size, model size, indicators per factor, and categories per item collectively when making analysis decisions.

Finally, Table 7 summarizes the issues previously summarized related to CFA with ordinal data, and the ways in which these issues are manifested in the data collected for the process of completion of CFA on the Evaluation Use and Evaluation Involvement Scales. These specific issues will feature prominently in decisions related to methodology in the CFA of these scales.

Table 7 *Specific Issues Related to the CFA of the Evaluation Use and Evaluation Involvement Scales*

Issue	*EU and EI Scale Data*
1. Sample size	Total N=300
2. Model size	Involvement Scales: 13 variables and 2 factors (based on EFA) Use Scales: 23 variables and 3 factors (based on EFA)
3. Number of categories per item	4: “No”, “Yes, a little”, “Yes, some”, “Yes, extensively”
4. Skewness Range	-0.55498 to 0.87231
5. Kurtosis Range	-1.72933 to -0.82611

CHAPTER 3 - METHOD

Introduction

The review of the literature provided an overview of the relevant research in the study of evaluation use and involvement that informed the creation of the Evaluation Use and Evaluation Involvement Scales. Also included was a review of literature describing the history of factor analysis as a data analytic technique and a discussion of the disagreements related to ordinal variables, particularly those measured using intensity response format items. This chapter will outline the methodology used for this study. It begins with a description of the research question, and also outlines the research design, sampling method, and process by which the issues related to CFA of this data were identified and addressed. The chapter concludes with a description of the ways in which these design elements allowed the researcher to address the research question guiding this study.

Research Question

The purpose of this research was to provide guidance to the evaluation community in the form of criteria that could be applied when making decisions about how best to implement CFA with ordinal survey data. The question driving this research was: *What are the data characteristics and analysis criteria that need to be considered to meet the assumptions of confirmatory factor analysis of ordinal evaluation survey data?* The data collected with the Evaluation Use and

Evaluation Involvement Scales was used as an example of the kinds of data that might be encountered by researchers and evaluators seeking construct validity evidence for the use of scales they have developed. Secondary to this research question was the potential provision of further validity evidence for the use of the Evaluation Involvement Scales and the Evaluation Use Scales in evaluations in the form of the results of the CFAs, to be described later in this chapter. This second focus of the research was less of a research question and more of a benefit to the field of evaluation resulting from the process of answering the primary question of this research project, and so it is embedded in the methodology designed to answer the main focus of the research.

Research Design

The process by which the research question was answered included four steps:

1. Articulation of the assumptions necessary to complete CFA, decisions related to completion of the analysis, and the characteristics of ideal data for analysis
2. Testing of the Evaluation Use and Evaluation Involvement Scales data against the assumptions of CFA, listed in step 1, and articulation of the characteristics of the data from the Scales
3. Completion of the CFAs of the two sets of scales, comparing the fits of the factors expected based on theory with the factors obtained through EFA

4. Summarization of the experience of completing CFA on the data obtained from the Evaluation Use and Evaluation Involvement Scales, including a list and description of guidelines for the application of CFA with similar data

Sampling Method

The survey data used to perform the CFAs are considered archival data in that they were collected during April and May, 2009. The data collected were designed to both provide summary data of the evaluation experience of the respondents for use by the National Science Foundation (NSF), who contracted the evaluations, and for potential use in an analysis that could provide further evidence of the validity of the Evaluation Use and Evaluation Involvement Scales. The process used to obtain the survey data is described to provide evidence of systematic sampling and also to provide evidence that the data to be used in the CFAs represents data from the same population as the data used in the two exploratory factor analyses that were conducted by other researchers, as mentioned in the review of the literature section.

Three NSF-funded programs with multi-site evaluations were originally selected by the Beyond Evaluation Use Grant (EUG) research team for inclusion in their study of evaluation use and involvement. The three programs were: (1) Advanced Technological Education Program (ATE); (2) Collaboratives for Excellence in Teacher Preparation (CETP); and (3) Local Systemic Change through Teacher Enhancement (LSC). These programs were purposefully

selected by the team for the study of use and involvement because the EUG researchers and staff members at the NSF felt that the programs represented a variety of levels of control in the evaluation process and therefore might represent different levels of the two main areas of interest in the research (Lawrenz & King, 2009). EFAs were then conducted on the items included in the survey given to the principal investigators (PIs)and co-principal investigators (co-PIs)of the projects throughout the United States that were funded as part of these three NSF programs.

The data used for the CFAs of the two sets of scales was collected in a similar manner to ensure that it came from the same identified population for comparison purposes. The population in this case was defined as PIs and co-PIs involved in project evaluations of large-scale, multi-site STEM programs funded by the NSF. The three programs selected for inclusion in the second survey implementation were: (1) Advanced Technological Education Program (ATE) – only those PIs and co-PIs whose projects had been funded after the first set of surveys was administered were included; (2) Math Science Partnership (MSP); and (3) Robert Noyce Teacher Scholarship Program (Noyce). Again, these programs were sampled to represent examples of large-scale, multi-site evaluations of STEM education programs funded by the NSF, though not purposefully along a continuum of expected involvement in the evaluation process as they were for the respondents whose data was used in the two EFAs. For the planned CFA, the team wanted to compare the fit of the resulting factors

from the EFAs with the factors originally developed based on theory. The two sampling plans were similarly designed to provide a comparison of samples from the same larger population of interest to the research team.

Table 8 *Survey Respondents by NSF program*

	ATE	*MSP*	*Noyce*	*Total*
Respondents	163	41	96	300

Names, institutional affiliations, and email addresses were obtained from the NSF website for each PI and co-PI on the projects funded under the three STEM education programs. Email notices of the survey, including a link to complete the online instrument, were sent to all PIs and co-PIs on the NSF lists for an attempt at a census of all those sampled from this population. First and second reminder emails were sent two and four weeks after the initial contact was sent and the survey was closed just under two months after the initial invitation to participate was made. The response rate for the survey was 44.8%. A non-response study was also conducted by the research team to determine whether the respondents were similar to the non-respondents to the survey. Though the team fell short on the goal of a census of all possible respondents, it was determined through a non-response study, described below, that the responses that were collected were fairly representative of all persons in the intended sample and therefore the population, though there is some evidence of potential response bias in the sample.

Non-response Study

To check for non-response bias, the research team conducted a non-respondent follow-up survey following the online survey of the PIs and co-PIs in each of the three programs. Non-response bias results when the people who responded to the survey are systematically different from the individuals in the sample who did not respond (Fowler, 2009). We developed a three-question non-respondent survey to address the topics of involvement and use, as well as to gather some information about what kept individuals from responding to the initial survey and reminder emails. Recipients of the non-respondent email were asked to select from the following responses: 1 – not at all involved; 2 – involved a little; 3 – involved some; and 4 – involved extensively. The non-respondent survey items were:

1. How involved were you in the *[Name of Program Evaluation]*?
2. How much impact did the *[Name of Program Evaluation]* have on you?
3. What was the main thing that kept you from responding to our initial survey request?

A random sample of non-respondents to the survey in each of the three programs (ATE, MSP, and Noyce) was drawn using Microsoft Excel. Each person was sent a personal, individually addressed email with a request for their input on the three items. Two reminder emails were sent to each of the individuals who did not respond. The table below shows the response rates of both the initial survey ("respondents") and the non-respondent follow-up survey ("non-respondents").

Table 9 *Response Rates to Original Survey and Non-response Survey*

Respondents		*Non-Respondents*	
45%	300/669	43%	60/139

The results of the analysis comparing the respondents and non-respondents showed differences in the overall levels of involvement and use reported. In each case the mean of the survey respondents was higher than the mean of the non-respondents. The Involvement Mean for the respondents was calculated as the mean of thirteen questions on the original survey. The Use/Impact Mean was calculated in the same way, using the fourteen survey questions related to evaluation use and impact. The means for the non-respondents were simply the means of their answers to the two questions they responded to via email. Though these measurements are not identical, the two sets of means closely mirror one another in that the 4-point scale for each question or set of questions is the same. The differences shown in Table 3 below may indicate the possibility of upward bias in the estimates of overall involvement and use/impact for survey respondents or it may also be an artifact of the difference between use, influence and overall impact. It is particularly difficult to separate the measurement of involvement in and use of an evaluation with response to a survey measuring these concepts, as it seems highly likely that a respondent's level of involvement would be related to his willingness to complete a survey. This fact makes a non-response study particularly difficult in this situation.

Table 10 *Summary of Findings of Mean Comparisons*

Involvement Mean		*Use/Impact Mean*	
Respondents	Non-respondents	Respondents	Non-respondents
2.28	1.54	2.40	1.53

Another way to determine the similarity of survey respondents and non-respondents is to compare them on known variables. To compare the respondents with the non-respondents (non-respondents in this comparison include all those who received the original invitation for the survey, but did not complete the survey, not just those who were sent the non-respondent email) the research team conducted chi square analyses of two demographic items that were available from the original project lists downloaded from the NSF website: role and project start date. Analyses were conducted to determine whether it was statistically significantly more likely that a respondent or non-respondent fit into different categories, including PI vs. co-PI and the year in which the projects began. As shown below in Tables 4 and 5, findings indicate a statistically significant finding for the respondent and role chi square analysis, but not for the respondent and project start year analysis. It appears that respondents who were listed as PIs for the projects selected for the study were more likely to be respondents to the survey and co-PIs were less likely to be respondents. This matches the anecdotal evidence that some of those who received the survey request emailed to state that someone else from their project had replied to the survey. Perhaps co-PIs assumed that the PI on the project would answer for the project. It was the intent of the study to have all potential respondents complete

the survey and, when emails such as those just described were received, this fact was communicated to potential respondents. Based on the second chi square analysis, it appears there is no statistically significant relationship between responding to the survey and the year in which an NSF-funded project began. These findings together show that there is possible response bias in the results as PIs were more likely to be respondents than were co-PIs and these two groups may respond differently to survey questions about evaluation use and involvement.

Table 11 *Chi Square Analysis: Respondent and Role*

	PI	*Co-PI*
Respondent	52%	40%
Non-respondent	48%	60%

Pearson χ^2 (1, n=669) = 8.545, p = .003, Cramer's φ = .113

Table 12 *Chi Square Analysis: Respondent and Project Start Year*

	2002	*2003*	*2004*	*2005*	*2006*	*2007*
Respondent	52%	45%	42%	41%	49%	43%
Non-respondent	48%	55%	58%	59%	51%	57%

Pearson χ^2 (1, n=664) = 3.592, p = .610, Cramer's φ = .074

Of the 60 responses to the non-respondent study, 48% reported that they did not feel the survey applied to them, and nearly 12% did not remember having received the initial survey. In addition, approximately 25% cited a lack of time. Over 15% offered a range of specific reasons in the "other" category, the basic themes of which were: a) they felt the survey was too long b) they felt the survey was unanswerable or didn't understand the questions; c) they were retired or

otherwise no longer involved in the project and d) a variety of individual explanations, such as family emergencies or sabbatical leaves. No one reported having had technical problems.

Methodological Process

Because previous researchers have already thoroughly described the process of development of the Evaluation Involvement and Evaluation Use Scales (Toal, 2007; Johnson, 2008), that description will not be included here. Rather, this section will include a description of the process by which the aforementioned research question was answered. Following is a description of how each of the four steps, outlined above, was completed.

1. Articulation of the assumptions necessary to complete CFA, decisions related to completion of the analysis, and the characteristics of ideal data for analysis

The assumptions necessary for completion of CFA include, 1) proper specification of the model, 2) measurement errors are independent of latent factors, and 3) measurement errors are uncorrelated with one another (Flora & Curran, 2004). Once a researcher determines that these basic assumptions are met and she can begin the analysis, decisions must be made related to the analysis based on the characteristics of the data being analyzed. Decisions must be made regarding: 1) method of parameter estimation, 2) measures of correlation, and 3) fit indices. Researchers of the CFA method have focused much effort on Monte Carlo simulations studying various characteristics of data and how these characteristics impact the results of the CFA. Some of these

characteristics include: minimum sample size of completed surveys, the size of the a priori model being tested, and the number of categories per item in the survey.

This first step in the research process articulated, in table form, the assumptions necessary to complete CFA, the decisions regarding the CFA that must be made, as well as when a researcher should make one decisions over another, and the characteristics of ideal data for completion of a CFA, based on the Monte Carlo simulation studies that have been published in the research literature on CFA. The MPlus software program was chosen for this set of CFAs due to the modification indices it provides when weighted least squares means and variance adjusted (WLSMV) is used as the parameter estimation method. This method appears to be best suited for use with ordinal data (Brown, 2006; Flora & Curran, 2004; Muthén & Muthén, 2007). Therefore the aforementioned table included information based on the MPlus CFA software.

2. Testing of the Evaluation Use and Evaluation Involvement Scales data against the assumptions of CFA, listed in step 1, and articulation of the characteristics of the data from the Scales

First, having articulated the basic assumptions necessary to complete a CFA in step 1, the data obtained from the conducting of the survey containing the Evaluation Use and Evaluation Involvement Scales was tested against these assumptions to ensure that completion of CFAs with this data was warranted. These assumptions were met. Next, the characteristics of the data relevant to

the decisions related to the CFA process listed in step 1 were articulated. Each of the three decisions was considered in turn, with decisions made as to, 1) which parameter estimation method to use, 2) which measure(s) of correlation to use, and 3) which fit indices to use in the analysis. Finally, the characteristics of the data from the two sets of scales were compared to the ideal CFA data characteristics listed in step 1. While the data was, unsurprisingly, not ideal, and the research is not conclusive as to specific guidelines for some of the data characteristics, a comparison between ideal data and actual data was conducted in order to consider how these differences might impact the results of the analysis.

3. *Completion of the CFAs of the two sets of scales, comparing the fits of the factors expected based on theory with the factors obtained through EFA*

Two different sets of confirmatory factor analyses were conducted; one of the Evaluation Use Scales and One of the Evaluation Involvement Scales. As mentioned previously, the MPlus software was used to conduct the analyses due to its inclusion of the WLSMV parameter estimation method, which has been shown to be successful with ordinal data. The adjustments determined to be necessary for these analysis based on the completion of steps 1 and 2 of the 4-step process was included in the CFAs conducted on the survey data. Because different factors were found in the EFAs conducted on the original set of data than were expected based on theory, the CFAs included a comparison of the models based on theory with the models found when the EFAs were conducted

on the data, using the comparative goodness of fit index (or indices) suggested by the research completed in steps 1 and 2. Tables 13 and 14 list the survey items and indicate those that were removed based on EFA results and research team discussion. Figures 5 through 8 follow the tables and outline the models used for the two sets of CFA analyses (Figures 5 and 6 for the CFAs comparing the theoretical and EFA models of the Evaluation Involvement Scales and Figures 7 and 8 for the CFAs comparing the theoretical and EFA models of the Evaluation Use Scales).

Table 13 *Survey Items Originally Included in the Evaluation Involvement Scales**

*Item #***	*Item Text*
1	I was involved in the discussions that focused the evaluation.
2	I was involved in identifying evaluation planning team members.
3	I was involved in developing the evaluation plan.
5	I was involved in developing data collection instruments.
6	I was involved in developing data collection processes.
7	I was involved in collecting data.
8	I was involved in reviewing collected data for accuracy and/or completeness.
9	I was involved in analyzing data.
10	I was involved in interpreting collected data.
12	I was involved in writing evaluation reports.
13	*I was involved in reviewing evaluation reports for accuracy and/or completeness.*
14	I was involved in presenting evaluation findings (e.g., to staff, to stakeholders, to an external audience).
15	*I was involved in developing future project plans based on evaluation results.*

*Note – Those items listed in italics are those that were removed based on EFA results and discussion by the research team that created the scales.
**Note – Item numbers that are missing are survey items that were not included in the scales. Item numbers were left as they were listed in the survey instrument to avoid confusion when relating those items that were removed based on the EFA results.

Table 14 *Survey Items Originally Included in the Evaluation Use Scales**

*Item #***	*Item Text*
25	The evaluation increased my knowledge/ understanding of how to plan an evaluation (e.g., discussing the focus of the evaluation, identifying evaluation planning team members, developing evaluation plan).
26	The evaluation increased my knowledge/ understanding of how to implement an evaluation (e.g., developing data collection instruments and processes, collecting, analyzing, reviewing, and interpreting data).
27	The evaluation increased my knowledge/ understanding of how to communicate evaluation findings (e.g., developing future plans for your project, writing and reviewing evaluation reports, and presenting evaluation findings).
28	The evaluation increased my knowledge /understanding of Science, Technology, Engineering, and Mathematics (STEM) education.
29	The evaluation increased my knowledge /understanding of STEM education evaluation.
30	The evaluation increased my knowledge/ understanding of my project.
31	The evaluation improved my skills in planning an evaluation (e.g., discussing the focus of the evaluation, identifying evaluation planning team members, developing the evaluation plan).
32	The evaluation improved my skills in implementing an evaluation (e.g., developing data collection instruments and processes, collecting, analyzing, reviewing, and interpreting data).
33	The evaluation improved my skills in communicating evaluation findings (e.g., writing and reviewing evaluation reports, presenting evaluation findings).
34	The evaluation improved my skills as a STEM educator.
35	The evaluation improved my skills as a STEM education evaluator.
36	*The evaluation improved my skills for working on my project.*
38	The evaluation increased my belief in the importance of planning an evaluation (e.g., discussing the focus of the evaluation, identifying

	evaluation planning team members, developing the evaluation plan).
39	The evaluation increased my belief in the importance of implementing an evaluation (e.g., developing data collection instruments and processes, collecting, analyzing, reviewing, and interpreting data).
40	The evaluation increased my belief in the importance of communicating evaluation findings (e.g., developing future plans for your project, writing and reviewing evaluation reports, presenting evaluation findings).
41	*The evaluation increased my belief in the importance of STEM education.*
42	The evaluation increased my belief in the importance of STEM education evaluation
43	The evaluation increased my belief in the importance of my project.
50	I used the program evaluation findings to make decisions about the future existence of my project (e.g., cancel or continue).
51	I used the program evaluation findings to develop future plans for my project.
52	I used the program evaluation findings to increase the attention given to my project.
53	I used the program evaluation findings to meet contractual and/or legal requirements of my project.
54	I used the program evaluation findings to enhance my organization's commitment to my project.
55	I used the program evaluation findings to seek funding for my project(s).
56	I used the program evaluation findings to promote additional evaluation of my project(s).
57	I used the program evaluation findings to make changes to my project(s).
58	I used the program evaluation findings to help plan a new project(s).

*Note – Those items listed in italics are those that were removed based on EFA results and discussion by the research team that created the scales.
**Note – Item numbers that are missing are survey items that were not included in the scales. Item numbers were left as they were listed in the survey instrument to avoid confusion when relating those items that were removed based on the EFA results. Items 45-49 were related to use in a future evaluation and were therefore only relevant to a subset of the population surveyed. The CFAs were conducted without these items in order not to limit the scales to those respondents who have been involved with another evaluation beyond the one considered in the survey.

Figure 5 Model of Evaluation Involvement (based on EFA results)

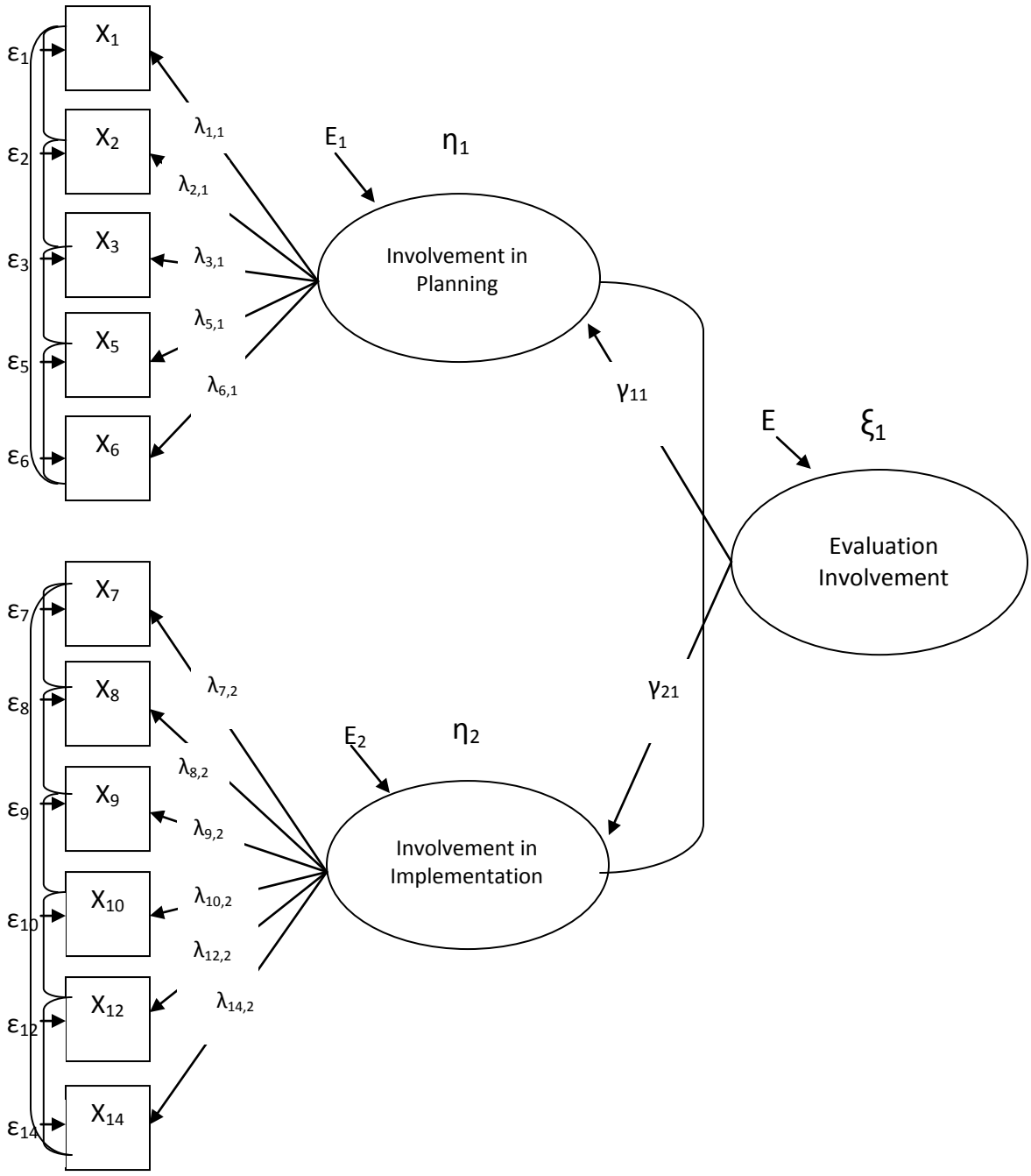


Figure 6 Model of Evaluation Involvement (based on theory)

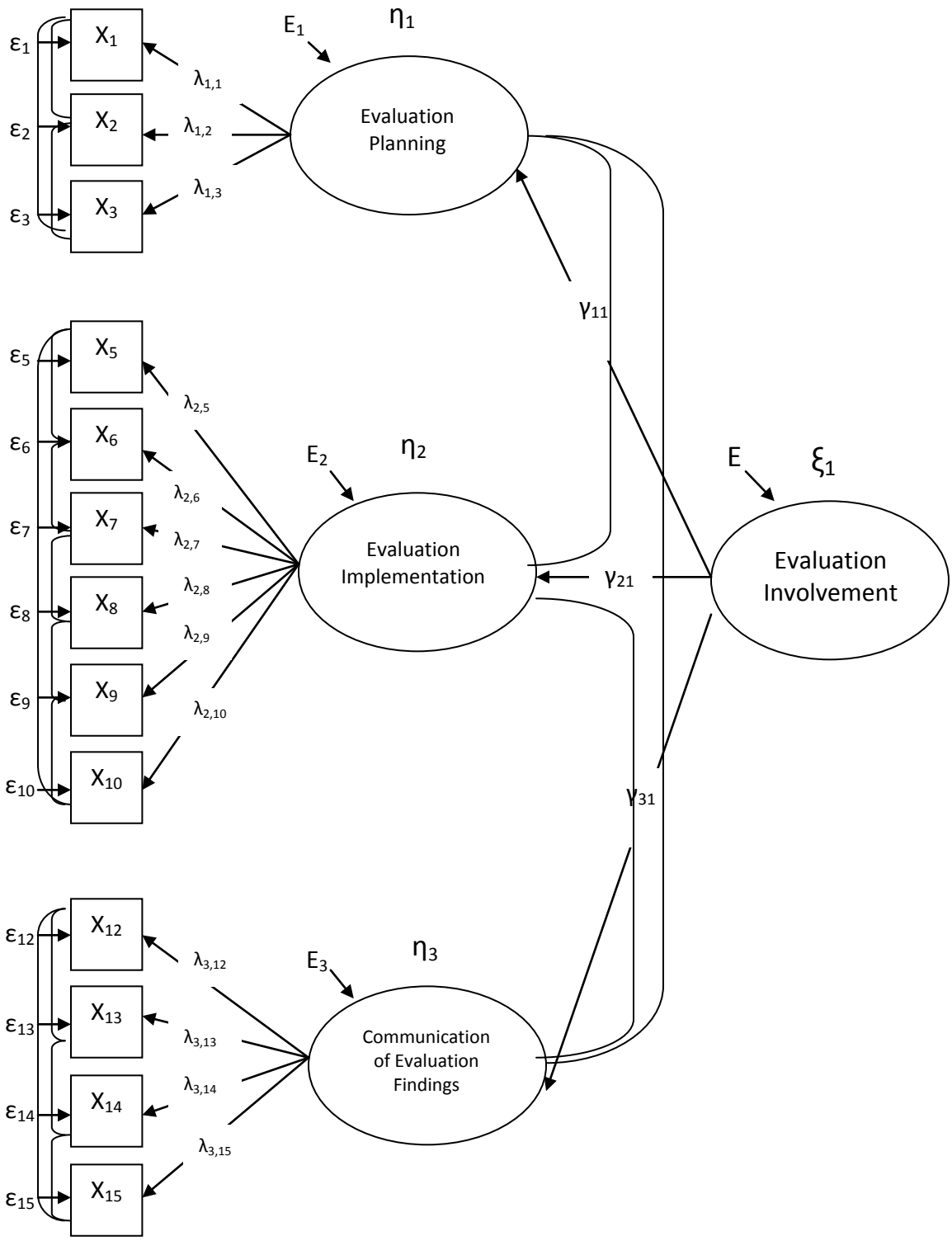


Figure 7 Model of Evaluation Use (based on EFA results)

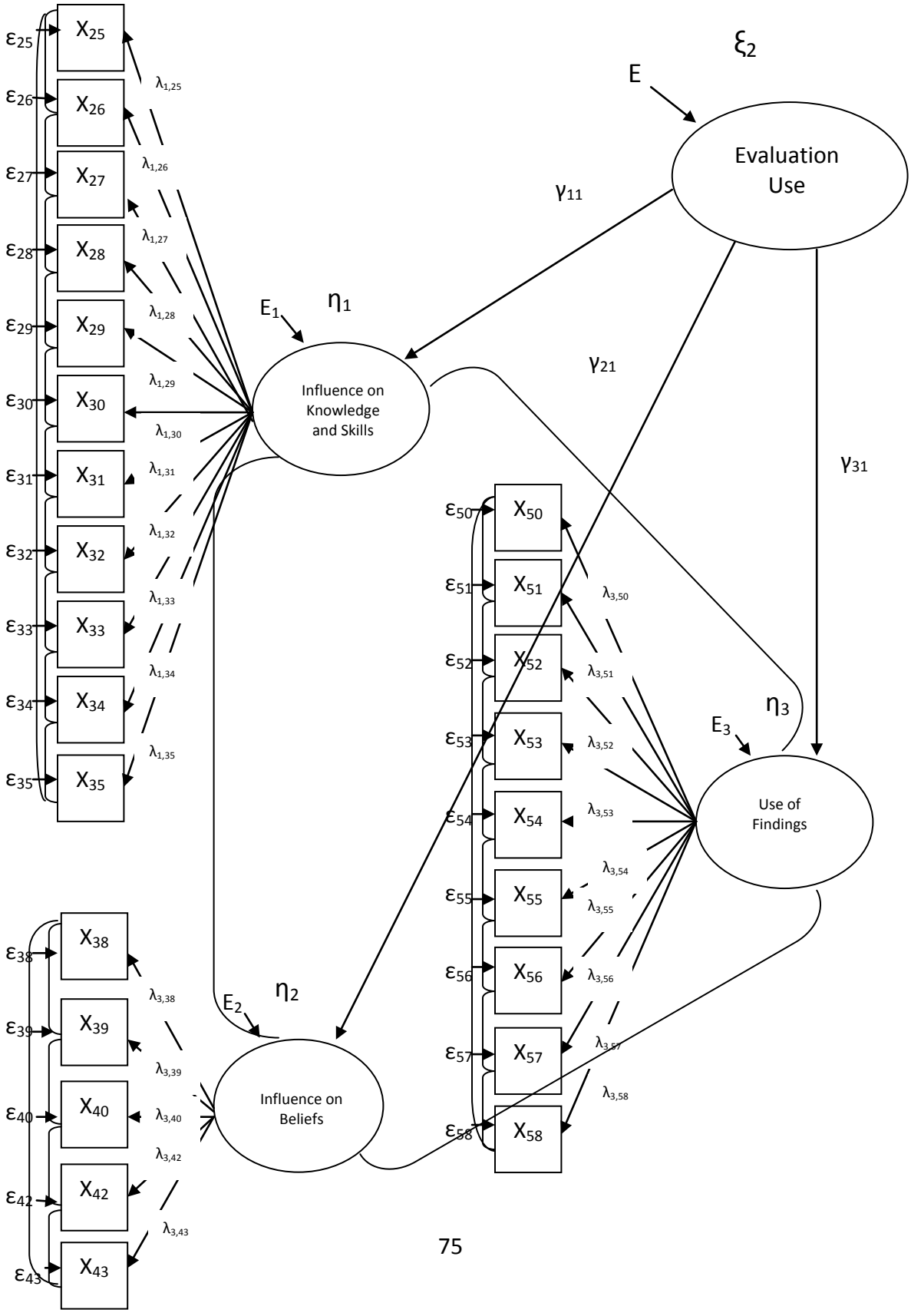
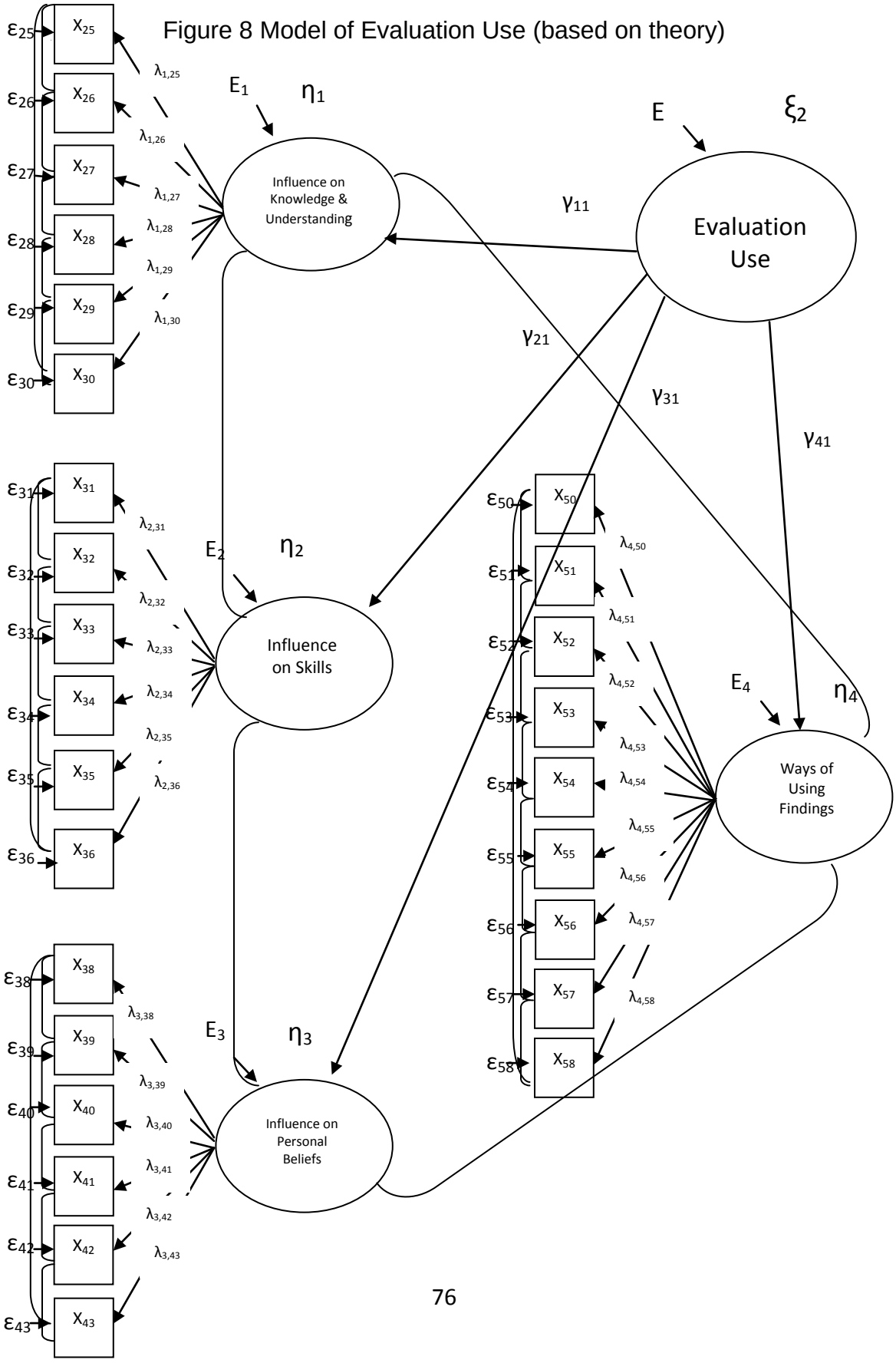


Figure 8 Model of Evaluation Use (based on theory)



4. Summarization of the experience of completing CFA on the data obtained from the Evaluation Use and Evaluation Involvement Scales, including a list and description of guidelines for the application of CFA with similar data

After completion of steps 1-3, two summaries were created. The first is a summary of the results of steps 1 and 2. This summary includes the articulation of the assumptions of CFA with ordinal data as well as a list of guidance from the literature on how to handle each characteristic of the Evaluation Use and Evaluation Involvement Scales data. Also included in this summary is a discussion of implications to interpretation of the CFA results resulting from the adaptations made to accommodate the data characteristics. Best practices, or ideal data characteristics, are included to inform those who may choose to use CFA with evaluation survey data of the characteristics their data should ideally achieve when planning a survey on which CFA will be performed. The second summary includes the results of the two CFAs conducted on each set of scales, including a statement about whether the results of the CFAs provide construct validity evidence for use of the two sets of scales with principal investigator and co-principal investigators of NSF-funded program evaluations. This summary models the best practices and guidelines included in the first summary, utilizing the extensive review of the CFA literature to inform the selection of adaptations made to the analysis and implications to the interpretation of the results.

Conclusion

The process outlined above was designed to enable the researcher to answer the question: *What are the data characteristics and analysis criteria that need to be considered to meet the assumptions of confirmatory factor analysis of ordinal evaluation survey data?* Factor analysis has been used as a tool to create, and provide validity evidence for the use of, surveys and other instruments of measurement since it was introduced by Spearman in 1904. The technique has been improved upon by experts in the fields of statistics and measurement since that time and it is now one of the most widely used techniques in the development of latent variable measures (Brown, 2006). Those who work in the field of evaluation might utilize this technique as they develop and provide validity evidence for the use of new measurement instruments that assess individual's use and involvement in evaluations, but only if the technique addresses the particular issues related to measurement instruments encountered by researchers and practitioners of evaluation.

Scales designed to measure both use and involvement in evaluation are needed, but it is difficult to provide validity evidence for instruments of this type because they often involve questions measured with intensity response formats which must be treated as ordinal data. The third step outlined above, the completion of CFAs of the Evaluation Use and Evaluation Involvement Scales, was conducted to potentially provide this needed validity evidence for use of these scales with persons involved in NSF-funded program evaluations. Steps 1,

2, and 4 were completed to assist evaluators and researchers who design and implement surveys in deciding whether CFA can successfully be completed with ordinal data with the characteristics of these two sets of scales.

CHAPTER 4 - RESULTS

Introduction

In the following chapter, the results of the four step process outlined in previous chapter will be considered in turn.

Step 1

Articulation of the assumptions necessary to complete CFA, decisions related to completion of the analysis, and the characteristics of ideal data for analysis

As listed in chapter 3, the assumptions for CFA include: 1) proper specification of the model, 2) measurement errors are independent of latent factors, and 3) measurement errors are uncorrelated with one another (Flora & Curran, 2004). The first two assumptions are not testable, but the third is, to some degree, tested during the completion of a CFA. The assumptions that should be tested prior to completion of a CFA are those associated with the parameter estimation being used. According to Brown (2006), multivariate normality of input indicators should be tested prior to completing a CFA using maximum likelihood parameter estimation. Given that this set of CFAs uses the weighted least squares parameter estimation method, testing of multivariate normality is unnecessary.

After considering CFA assumptions, it was important to list the decisions related to selecting components of the analysis where choices are available, as

well as the default settings, or options available, in the MPlus software for each of these choices. Table 15 lists these choices and default settings.

Table 15 *CFA Analysis Choices and MPlus Default Settings*

CFA Component Decision	*MPlus Default/Options*
Parameter estimation method	ML is most commonly used and is the default option with continuous factor indicators. WLSMV is default method with categorical factor indicators, and the syntax needs to include the fact that the data are categorical (Muthén & Muthén, 1998-2007)
Measure of correlation	Polychoric is the default measure with categorical data having more than 2 options per factor indicator (Muthén & Muthén, 1998-2007)
Measures of fit	Goodness of fit indices provided in Mplus with WLSMV parameter estimation include: chi square, CFI, TLI, AIC, BIC, adjusted BIC, RMSEA, WRMR (Muthén & Muthén, 1998-2007)

After considering the assumptions of CFA and the analysis choices available, another consideration to make prior to completion of the CFA is a thorough review of the categories of data characteristics that may impact the results of the analysis. Statistical analyses are often created with ideal data characteristics in mind. Initial Monte Carlo simulation studies also often use data simulated to fit ideal characteristics rather than characteristics reflective of data found in everyday usage. Though data is not expected to be ideal, it is important to consider what the data used in the analysis is expected to be like and how differing from that ideal might impact the results of the CFA. Table 16 summarizes the data characteristic categories considered, the ideal data

characteristics for completion of CFA, and how using data that differs from this ideal might potentially impact the results of the analysis.

Table 16 *Consideration of Ideal CFA Data Characteristics*

Data characteristic	*Ideal data description*	*Implications for non-ideal*
Sample size	There is no agreed-upon minimum sample size, especially when the level of communalities, number of factors, and number of indicators per factor can be different for each analysis (MacCallum, Widaman, Zhang, & Hong, 1999). It is recommended that a Monte Carlo simulation be completed in order to determine the minimum sample size needed for each model in order to obtain power of .8 (Muthén & Muthén, 2002; Brown, 2006)	If the sample size is not large enough, power will be negatively impacted (Muthén & Muthén, 2002); As sample size increases, factor loadings will have smaller standard errors. Sample size can also affect factor solutions as solutions obtained from larger samples have been found to be more stable and to more accurately recover the population loadings (MacCallum, Widaman, Zhang, & Hong, 1999). WLS has shown biased parameter estimates with small samples, but WLSMV is less susceptible to sample size issues (DiStefano, 2002; Flora & Curran, 2004).
Model complexity	There really is no ideal level of model complexity, but there is an interaction between the complexity of a model (e.g. number of factors, level of communality, and level of overdetermination) and the sample size needed to have confidence in the results (MacCallum, Widaman, Zhang, & Hong, 1999).	The implication for having a complex model is that it likely increases the number of samples necessary to achieve a power value of .8, which allows the researcher to have confidence in the results of the analysis. (MacCallum, Widaman, Zhang, & Hong, 1999).

Categories per item	Dolan (1994) recommends a minimum of 5 categories per item to effectively estimate the chi square value.	Having fewer than 5 categories per item may impact the reliability of the chi square value, but this may not have a large impact on the analysis overall as it is not recommended to use the chi square as an index of fit, which will be discussed later.

Once the assumptions necessary to complete CFA, decisions related to completion of the analysis, and the characteristics of ideal data for analysis were articulated as well as the potential implications of using data that is less than ideal were considered, the Evaluation Involvement and Evaluation Use scale data was considered with these results in mind.

Step 2

Testing of the Evaluation Use and Evaluation Involvement Scales data against the assumptions of CFA, listed in step 1, and articulation of the characteristics of the data from the Scales

As mentioned in step 1, above, it is not necessary to test the data against any assumptions prior to completion of a CFA. While the maximum likelihood parameter estimation method does have associated assumptions that are testable, the weighted least squares estimation method does not. Therefore this section will include only the articulation of the characteristics of the data collected using the Evaluation Use and Evaluation Involvement Scales as it relates to the decisions and implications outlined in step 1.

Table 17 *CFA Analysis Choices and Selection for Analyses*

CFA component decision	*Recommendations*	*Selection for these CFAs*
Parameter estimation method	WLSMV is recommended for use with categorical data (Flora & Curran, 2004; Brown, 2006; Muthén & Muthén, 2007)	WLSMV (robust weighted least squares)
Measure of correlation	Polychoric correlation is recommended for use with WLS estimation method (Babakus, Ferguson, & Jöreskog, 1987; DiStefano, 2002; Flora & Curran, 2004)	Polychoric correlation
Measures of fit	Recommends considering 1 of each of 3 types: absolute, parsimony correction, and comparative fit (Brown, 2006). Specifically recommend: Absolute fit: chi square with its df and p-value (despite its problems), RMSEA (which Brown considers parsimony correction), SRMR; Incremental fit (which seems to be the same as Brown's comparative): CFI; Parsimony fit: PNFI (Hooper, Coughlan, & Mullen, 2008). Recommend RMSEA with analysis of categorical data (Hutchinson & Olmos, 1998).	Absolute Fit: Report chi square value along with df and p-value, but do not use this information to determine fit; use WRMR as fit index for this category (because it is offered in MPlus instead of SRMR with WLSMV estimation); Parsimony Correction: Report RMSEA; Comparative Fit: Report CFI.

	Recommends CFI for comparative fit index used with analysis of categorical data (Bentler, 1990).	

Based on recommendations in the measurement literature, summarized both in Chapter 2 and Table 17, the CFAs conducted in this study used the robust least squares parameter estimation method (WLSMV) with polychoric correlation. The goodness of fit indices that were considered when assessing the fit of the models based on the data included: weighted root mean square residual (WRMR); root mean square error of approximation (RMSEA); and comparative fit index (CFI). The chi square value was also reported for each CFA but, since it is highly sensitive to sample size and almost always significant with large samples, it was not considered as a measure of goodness of fit (Harrington, 2009).

The three considerations listed in Table 16 may have implications for the outcome of the CFAs, which will be discussed later, but decisions were not made related to these considerations as the sample size, model complexity, and categories per item were already set at the start of this set of analyses. The Monte Carlo sample size simulation mentioned in Table 16 was conducted after the CFAs to assess whether the sample size of 300 was large enough to achieve an adequate power value in the analysis. The results of the simulations conducted will be presented in step 3, below.

Step 3

Completion of the CFAs of the two sets of scales, comparing the fits of the factors expected based on theory with the factors obtained through EFA

To report the results of the two CFAs performed on the Evaluation Use Scale data (one model based on theory and one model based on the results of an EFA) and the two CFAs performed on the Evaluation Involvement Scale data (one model based on theory and one model based on the results of an EFA) a summary of results for all four CFAs is presented below in a structure based on the recommendation of Brown (2006).

I. Model Specification
    a. Conceptual/empirical justification for models (in literature review)
    b. Description of the parameter specification of the model (see models in chapter 3)
    c. Demonstration that the model is identified: The latent factors in each of the four CFA models were permitted to be correlated as the theory behind the scale development suggested that the factors were related. The models were all overidentified, with the following degrees of freedom:
        i. Involvement from Theory: 25 *df*
        ii. Involvement from EFA: 16 *df*
        iii. Use from Theory: 50 *df*
        iv. Use from EFA: 51*df*

II. Input Data

a. Description of sample characteristics, size, and sampling method (see description in chapter 3)
b. Description of the type of data used (see full description in chapter 3):
    i. Ordinal data on a 4-point intensity format of 1=No; 2=Yes, a little; 3=Yes, some; 4=Yes, extensively
c. Test of estimator assumptions

While the use of maximum likelihood (ML) parameter estimation necessitates the assessment of the data for multivariate normality, robust weighted least squares (WLSMV), the parameter estimation methods used in these CFA analyses does not require that the data be multivariate normal.

d. Missing data
    i. Extent and nature:
        1. Evaluation Involvement Scales: 4/300 had no data for these items, reducing the sample size to 296. Of these 296, 38 survey respondents had missing data on between 1 and 4 items.
        2. Evaluation Use Scales: 4/300 had no data for these items, reducing the sample size to 296. Of these 296, 31 survey respondents had missing data on between 1 and 3 items.

ii. Method of missing data management: MPlus uses pairwise deletion as default (Muthén & Muthén, 1998-2007), and this is the best option in this case due to the small number of missing values in this dataset and the fact that the use of pairwise deletion means the sample size is not unnecessarily reduced, especially given the fact that so little data was missing for each respondent.

III. Model Estimation

a. Software version used: MPlus version 5.21

b. Type of data/matrices analyzed: polychoric correlations

c. Estimator used: robust weighted least squares (WLSMV), (explanation for selection of this estimator is included in Step 2).

IV. Model Evaluation

a. Overall goodness of fit:

Table 18 *Overall Goodness of Fit Results*

	Involvement from Theory	*Involvement from EFA*	*Use from Theory*	*Use from EFA*
Absolute Fit:				
chi square value	191.232	175.171	367.465	336.762
df	25	16	50	51
p-value	<.001	<.001	<.001	<.001
WRMR value	1.309	1.577	1.427	1.345
Parsimony Correction:				
RMSEA value	0.150	0.183	0.146	0.138
Comparative Fit:				
CFI value	0.958	0.962	0.946	0.953

The cutoff values that should be used for the goodness of fit indices reported above include the following: WRMR ≤ 1.0 (Yu, 2002), RMSEA ≤ .06,

and CFI ≥ .95 (Hu & Bentler, 1999). Using these recommended cutoff values for assessing goodness of fit for these CFAs, only the CFI values for both of the Involvement models and the Use EFA model meet the cutoff criteria. There is little evidence for good model fit in these analyses.

b. Localized areas of ill fit:

i. Standardized Residuals

Table 19 *Standardized Residual Ranges for Each CFA*

Involvement from Theory	*Involvement from EFA*	*Use from Theory*	*Use from EFA*
-0.150 to 0.141	-0.191 to 0.117	-0.150 to 0.132	-0.128 to 0.136

Residuals of concern are those equal to or greater than the absolute value of 1.96 (Brown, 2006). None of the residuals in any of the four CFAs fit within this range of concern, providing little evidence of localized areas of ill fit in the models.

ii. Modification Indices

Table 20 *Modification Indices Greater than 4.0*

Involvement from Theory			*Involvement from EFA*			*Use from Theory*			*Use from EFA*		
Item	MI	EPC	Item	MI	EPC	Item	MI	EPC	Item	MI	EPC
f1 by 5	7.3	0.23	f1 by 8	4.2	-0.24	f1 by 36	8.6	0.78	f1 by 38	4.2	-0.12
f1 by 6	8.3	0.24	f1 by 9	4.7	-0.25	f1 by 58	7.4	0.17	f1 by 40	4.8	-0.13
f1 by 7	11.8	-0.42	f1 by 14	21.6	0.55	f2 by 25	4.0	-0.49	f1 by 42	6.3	0.16
f1 by 8	7.3	-0.30	f2 by 1	7.7	-0.30	f2 by 58	7.2	0.16	f1 by 58	10.0	0.17
f1 by 12	12.1	-0.56	f2 by 2	4.4	-0.23	f3 by 25	4.9	-0.17	f2 by 30	5.1	0.20
f1 by 15	14.2	0.50	f2 by 3	7.1	-0.28	f3 by 31	9.1	-0.24	f2 by 31	7.2	-0.19
f2 by 12	19.5	0.95	f2 by 6	26.2	0.52	f3 by 34	4.2	0.17	f2 by 58	9.2	0.18
f2 by 15	15.1	-0.79				f3 by 36	5.2	0.17	f3 by 30	35.3	0.35
f3 by 7	4.3	-0.50				f3 by 58	6.3	0.16	f3 by 31	12.6	-0.21
						f4 by 30	15.9	0.27	f3 by 32	5.6	-0.14
						f4 by 31	10.1	-0.21	f3 by 38	8.9	-0.17
						f4 by 32	5.2	-0.15	f3 by 43	16.7	0.25
						f4 by 36	19.9	0.27			
						f4 by 38	5.6	-0.14			
						f4 by 43	8.8	0.19			

MI = modification index value; EPC = (completely standardized) expected parameter change value (how much the parameter is expected to change, positively or negatively, if freely estimated in a subsequent analysis (Brown, 2006).

Modification indices of 3.84 (rounded up to 4.0 in practice) or greater suggest that model overall fit could be improved if the fixed parameter was freely estimated (Brown, 2006). The items listed in Table 20, above, had modification indices greater than 4.0. Since freely estimating parameters should only be done when doing so is supported by empirical, conceptual, or practical considerations (Brown, 2006) and there is not enough research on these scales at this point to support the free estimation of parameters, these modification indices are presented as evidence of the existence of localized areas of ill fit. No changes to the estimation of parameters are suggested at this early point in the development of the Evaluation Involvement and Evaluation Use Scales. However, the fact that the modification indices show some relatively high crossloadings of items with factors they are not predicted to be related to provides further evidence of localized areas of ill fit and should be noted in future analyses of these scales.

c. Other sources of poor fit

Brown (2006) suggests a number of sources of poor-fitting CFA solutions, and two that may be relevant to these analyses are summarized below.

i. Problematic discriminant validity

When too many factors have been specified in the model, this issue can often be detected by high correlations between factors (Brown, 2006). Correlations above .85 are generally considered to indicate potential problems with discriminant validity. Table 21 lists the factor correlations for the four CFAs.

Table 21 *Factor Correlations*

CFA Model	*Correlations*
Involvement from Theory	Planning with Implementation: .640 Planning with Communication: .762 Implementation with Communication: .845
Involvement from EFA	Planning with Implementation: .726
Use from Theory	Knowledge with Skills: .918 Knowledge with Influence on Beliefs: .643 Knowledge with Use of Findings: .576 Skills with Influence on Beliefs: .661 Skills with Use of Findings: .543 Influence on Beliefs with Use of Findings: .498
Use from EFA	Knowledge & Skills with Influence on Beliefs: .671 Knowledge & Skills with Use of Findings: .558 Influence on Beliefs with Use of Findings: .498

The two factor correlations of potential concern are Implementation with Communication (r = .845) and Knowledge with Skills (r = .918). This finding is not surprising given the fact that both sets of EFAs resulted in one fewer factor than was expected based on theory. In the case of the Evaluation Use Scales, Knowledge and Skills, the two highly correlated factors, results in one factor, Knowledge & Skills, in the EFA.

Though none of the factor correlations in the two CFA models based on EFA results were larger than the recommended cutoff of .85, the high correlations among factors may have an impact on poor model fit.

ii. Incorrect designation of the relationships between indicators and latent factors

Relationships between indicators and latent factors can manifest in three ways: 1) an indicator can load on more than one factor when specified in the model to load on only one, 2) an indicator can be specified to load on the incorrect factor, or 3) an indicator can be specified to load on a factor when it actually has no relationship to any factors in the model (Brown, 2006). Modification indices may indicate items (indicators) that should load on more than one factor. Consideration of the modification indices listed in Table 20 show items that may load on multiple factors, particularly in the Evaluation Use Scales. Items 7 and 15 have high modification indices on multiple factors in the Evaluation Involvement models, though item 15 was removed as a result of the EFA and therefore does not appear in that model.

In the Evaluation Use models, items 25, 30, 31, 36, 38, and 58 have modification indices on multiple factors (item 36 was removed as a result of the EFA and does not appear in that model). In fact, item 30 is the item with the largest modification index (35.29) and, based on general recommendations, would be the most likely candidate for the freeing of its fixed parameter. Given the modification indices resulting from these analyses, it is possible that the relationships between indicators and latent factors are incorrectly designated which could lead to poor model fit.

To summarize the potential sources of the poor model fit resulting from the CFAs conducted on the four models, it would appear that both the relatively large correlations between factors and large number of items with modification indices

above 4.0 may negatively impact the fit of the models. There is not yet sufficient reason to freely estimate fixed parameters, such as those items with the highest modification indices, but it is clear that these areas are of particular concern to the fit of these models.

V. Parameter Estimates

Table 22 *Parameter Estimates for Involvement from Theory*

Factor	*Item*	*λ*				R^2	*ε*	*Φ*
		Unstandardized	*S.E.*	*Standardized**	*S.E.*			
1	1	1.000	0.000	0.942	0.013	0.887	0.113	1.00
	2	1.002	0.017	0.944	0.013	0.891	0.109	
	3	1.017	0.020	0.957	0.012	0.916	0.084	
2	5	1.000	0.000	0.913	0.016	0.834	0.166	1.00
	6	1.048	0.023	0.957	0.011	0.916	0.084	
	7	0.789	0.038	0.720	0.035	0.519	0.481	
	8	0.926	0.026	0.846	0.023	0.715	0.285	
	9	1.028	0.021	0.938	0.012	0.880	0.120	
	10	1.011	0.022	0.923	0.013	0.852	0.148	
3	12	1.000	0.000	0.860	0.027	0.739	0.261	1.00
	13	1.052	0.033	0.905	0.016	0.818	0.182	
	14	1.037	0.035	0.892	0.020	0.795	0.205	
	15	1.038	0.039	0.893	0.021	0.797	0.203	

*All estimates are statistically significant (p<.001). All estimates are standardized with the exception of the unstandardized factor loadings, which are labeled.

Factor correlations: F1 with F2 = 0.640, F1 with F3 = 0.762, F2 with F3 = 0.845

Table 23 *CFA Results for Involvement from EFA*

Factor	*Item*	*λ*				R^2	*ε*	*Φ*
		Unstandardized	*S.E.*	*Standardized**	*S.E.*			
1	1	1.000	0.000	0.924	0.013	0.855	0.145	1.000
	2	0.998	0.018	0.922	0.014	0.851	0.149	
	3	0.998	0.019	0.923	0.014	0.852	0.148	
	5	0.989	0.020	0.914	0.016	0.836	0.164	
	6	1.040	0.019	0.962	0.011	0.925	0.075	
2	7	1.000	0.000	0.741	0.034	0.549	0.451	1.000
	8	1.151	0.049	0.853	0.022	0.728	0.272	
	9	1.290	0.057	0.956	0.011	0.914	0.086	
	10	1.245	0.055	0.923	0.013	0.852	0.148	
	12	1.154	0.054	0.855	0.026	0.731	0.269	
	14	1.177	0.057	0.872	0.025	0.761	0.239	

*All estimates are statistically significant (p<.001). All estimates are standardized with the exception of the unstandardized factor loadings, which are labeled.

Factor correlation: F1 with F2 = 0.726

Table 24 *CFA Results for Use from Theory*

Factor	*Item*	*λ*				R^2	*ε*	*Φ*
		Unstandardized	*S.E.*	*Standardized**	*S.E.*			
1	25	1.000	0.000	0.956	0.006	0.914	0.086	1.000
	26	0.999	0.009	0.955	0.007	0.912	0.088	
	27	0.989	0.008	0.946	0.008	0.894	0.106	
	28	0.928	0.016	0.888	0.016	0.788	0.212	
	29	0.962	0.010	0.920	0.010	0.846	0.154	
	30	0.905	0.018	0.866	0.017	0.749	0.251	
2	31	1.000	0.000	0.965	0.006	0.931	0.069	1.000
	32	0.994	0.009	0.959	0.007	0.919	0.081	
	33	0.976	0.009	0.942	0.009	0.887	0.133	
	34	0.905	0.017	0.873	0.017	0.763	0.237	
	35	0.958	0.011	0.925	0.011	0.855	0.145	
	36	0.959	0.012	0.925	0.011	0.856	0.144	
3	38	1.000	0.000	0.953	0.007	0.909	0.091	1.000
	39	1.010	0.009	0.963	0.007	0.927	0.073	
	40	1.013	0.009	0.966	0.006	0.933	0.067	
	41	0.919	0.020	0.876	0.018	0.767	0.233	
	42	0.985	0.012	0.939	0.010	0.882	0.118	
	43	0.903	0.023	0.861	0.021	0.742	0.258	
4	50	1.000	0.000	0.893	0.015	0.798	0.202	1.000
	51	1.066	0.023	0.953	0.014	0.907	0.093	
	52	1.003	0.023	0.896	0.017	0.803	0.197	
	53	0.789	0.040	0.705	0.036	0.497	0.503	
	54	0.957	0.023	0.855	0.020	0.731	0.269	
	55	0.967	0.024	0.864	0.020	0.746	0.254	
	56	0.895	0.033	0.800	0.028	0.640	0.360	
	57	1.001	0.019	0.894	0.017	0.800	0.200	
	58	0.998	0.023	0.892	0.020	0.795	0.205	

*All estimates are statistically significant (p<.001). All estimates are standardized with the exception of the unstandardized factor loadings, which are labeled.

Factor correlations: F1 with F2 = 0.918, F1 with F3 = 0.643, F1 with F4 = 0.576, F2 with F3 = 0.661, F2 with F4 = 0.543, F3 with F4 = 0.498.

Table 25 *CFA Results for Use from EFA*

Factor	*Item*	*λ*				R^2	*ε*	*Φ*
		Unstandardized	*S.E.*	*Standardized**	*S.E.*			
1	25	1.000	0.000	0.953	0.007	0.909	0.091	1.000
	26	0.995	0.008	0.949	0.007	0.901	0.099	
	27	0.985	0.008	0.939	0.008	0.833	0.117	
	28	0.921	0.016	0.878	0.016	0.772	0.228	
	29	0.955	0.010	0.911	0.011	0.830	0.170	
	30	0.880	0.020	0.839	0.020	0.704	0.296	
	31	1.008	0.008	0.961	0.007	0.924	0.076	

	32	1.002	0.009	0.955	0.007	0.913	0.087	
	33	0.980	0.010	0.935	0.009	0.874	0.126	
	34	0.896	0.019	0.854	0.019	0.729	0.271	
	35	0.961	0.012	0.916	0.012	0.840	0.160	
2	38	1.000	0.000	0.955	0.007	0.913	0.087	1.000
	39	1.010	0.009	0.965	0.007	0.931	0.069	
	40	1.014	0.008	0.968	0.006	0.938	0.062	
	42	0.967	0.014	0.923	0.013	0.853	0.147	
	43	0.883	0.025	0.844	0.024	0.712	0.288	
3	50	1.000	0.000	0.897	0.015	0.804	0.196	1.000
	51	1.060	0.022	0.950	0.014	0.903	0.097	
	52	1.000	0.022	0.896	0.017	0.803	0.197	
	53	0.787	0.039	0.705	0.036	0.498	0.502	
	54	0.949	0.023	0.851	0.021	0.724	0.276	
	55	0.966	0.023	0.866	0.020	0.751	0.249	
	56	0.891	0.032	0.799	0.028	0.639	0.361	
	57	0.997	0.019	0.894	0.017	0.799	0.201	
	58	0.996	0.023	0.893	0.019	0.797	0.203	

*All estimates are statistically significant (p<.001). All estimates are standardized with the exception of the unstandardized factor loadings, which are labeled.

Factor correlations: F1 with F2 = 0.671, F1 with F3 = 0.558, F2 with F3 = 0.498.

All factor loadings in each of the four models are > .71, with the exception of item 53 in the Use models (λ = .705). Based on general rules of thumb, factor loadings this large are considered excellent (Harrington, 2009).

a. Verification of power and precision of the model estimates

In order to assess the power of the analysis, a Monte Carlo simulation was conducted for each of the four CFAs, based on the sample size simulations suggested by Muthén and Muthén (2002) and illustrated by Brown (2006). For each simulation, parameter estimates, residual variances, and thresholds were obtained from the CFAs conducted on each model. Four criteria are used to determine the appropriate sample size for a CFA (Brown, 2006):

1. The bias of the parameters and their standard errors do not exceed 10% for any parameter in the model.

2. For parameters that are the specific focus of the power analysis (covariances of the factors), the bias of their standard errors should not exceed 5%.

3. Coverage for all parameters should range between .91 and .98.

4. The power of the covariance between factors should be ≥ .80.

Table 26 *Results of Monte Carlo Sample Size Simulations*

Model	*Criteria 1 (Range)*	*Criteria 2 (Range)*	*Criteria 3 (Range)*	*Criteria 4 (Range)*
Involvement (Theory)				
Involvement (EFA)				
Use (Theory)				
Use (EFA)				

*Note: Sample size was set to 296 for each simulation as that was the sample size obtained for the study; Bold values indicate criteria is met, suggesting the sample size of 296 is adequate to obtain necessary power.

Results suggest…

VI. Substantive Conclusions

The following summary is included in this section rather than chapter 5 so that a complete example of written results of a CFA can be provided in its entirety. Though two other researchers found support for two factors and three factors in the Evaluation Involvement and Evaluation Use Scales, respectively, the completion of CFAs on data from a separate administration of the scales did not provide construct validity evidence for the factoring of the items in the same manner. Analyses of the scale items as they were projected to load on factors

based on theory also did not prove fruitful in terms of searching for construct validity evidence in support of factoring the items in this way.

Both the factors predicted by theory when creating the two sets of scales and the resulting factors from conducting EFAs on data collected with the scales were tested using CFAs in order to search for potential differences between the models. Even if the CFAs would have resulted in goodness of fit indices within the recommended cutoffs, there was no way to statistically test the results to find out which model is "better", as the fit indices are used to assess fit, not compare fitness of different models. Had one or the other model met the cutoff criterion while the other didn't, an argument might have been made that one model had evidence of proper fit while the other did not, but the results provided little evidence of proper fit for any of the models.

*Add statement about sample size if found to be too small in MC simulation

Another area of particular concern in the results of these CFAs is the modification indices, summarized in Table 20, above. A large number of items had modification indices above 4.0, indicating the possible need to freely estimate some fixed parameters. With freely estimated parameters, the CFA solution parameters that optimally reproduce the variances and covariances of the input matrix are found in the process of the analysis. With fixed parameters, the researcher assigns values based on a priori theory (Brown, 2006). In order to justify free estimation of parameters, there should be theoretically based

decisions to do so. Given the small amount of research that has been done on these scales, however, it was not advisable at this time to make decisions about which parameters to estimate freely. More research should be conducted in order to make changes based on these modification indices findings.

Step 4

Summarization of the experience of completing CFA on the data obtained from the Evaluation Use and Evaluation Involvement Scales, including a list and description of guidelines for the application of CFA with similar data (guidelines provided in chapter 5)

A large body of literature was referenced for the creation of the Evaluation Use and Evaluation Involvement Scales. The creation of these scales marked the first large-scale attempt to measure use and involvement in large, multi-site evaluations using a survey. EFAs conducted on the two sets of scales resulted in a factor structure quite similar to that expected based on the theoretical bases for the scales.

In chapter 2, the disagreements related to the use of ordinal data were presented. One of these issues was the historical disagreement regarding which statistical analyses are appropriately conducted on ordinal data, with some consensus indicating that only means should be produced, and no advanced analyses, such as regression analysis, advised. This fact necessitates the provision of factors to be used in further analyses of the data, which means factor analyses should be conducted to provide validity evidence that the individual

items on the survey load on the factors as expected by theory. CFA can be used to provide this validity evidence.

Issues related to intensity response format items were also presented in chapter 2. These items are most frequently treated as ordinal data and have been successfully used in factor analyses. While some literature suggests 5 categories per item is optimum for completion of CFA on the data, analyses have been conducted on data with 2, 3, and 4 categories. The Evaluation Use and Evaluation Involvement scales have 4 categories per item and, had they been designed with 5 categories due to the potential optimization for the conducting of CFAs on the data, it is not known how this change would impact the responses of those who completed the survey containing the scales. It is common practice for CFAs to be completed on ordinal data with fewer than 5 categories per item, and the major software programs used for CFA allow for ordinal data with fewer than 5 categories, therefore, having data with 4 categories did not appear to be an issue.

Considerations in completing CFA with ordinal data included: which software best adapts to ordinal data, which parameter estimation method to use, which measure of correlation to use, and which goodness-of-fit indices were most accurate when using ordinal data in the analysis. Though an attempt was made to compare the model based on theory for each scale with the results of the EFAs conducted on each scale, the literature suggests that goodness-of-fit indices should not be used to decide which model is a better fitting model, as

they are designed to act as cutoff criteria to provide evidence of fit. Therefore it is not advisable to use them to compare models.

CHAPTER 5 - DISCUSSION AND IMPLICATIONS

The research question guiding this study was: What are the data characteristics and analysis criteria that need to be considered to meet the assumptions of confirmatory factor analysis of ordinal evaluation survey data? Following is a summary discussion of the results of the analysis pertaining to the research question, including a consideration of the data characteristics and analysis criteria related to completion of CFA. The summary includes lessons learned while completing the CFAs of the Evaluation Use and Evaluation Involvement Scales and guidelines for evaluators considering the use of CFA in the process of designing a survey.

Summary of Findings

Data Characteristics

1. Sample size

A sufficiently large sample size is needed in any CFA to achieve adequate power to trust the analysis results. Though rules of thumb relating to sample size have been suggested, a more specific method of determining sample size should be completed. One such method suggested by multiple sources is a Monte Carlo simulation in which the sample size is adjusted to determine the size needed to achieve power of .8, considered adequate for trusting the results of the analysis (Muthén & Muthén, 2002; Brown, 2006; Harrington, 2009).

Sample size as a data characteristic is specifically related to CFA with ordinal data in that the parameter estimation method recommended for use with this

data type, WLSMV (summarized below) , is particularly sensitive to sample size. Conducting a CFA with ordinal data requires the use of this estimation method, which can require a larger sample size than do other methods of estimating parameters. It is therefore of particular concern that an adequate sample size is possible to achieve when a CFA is being considered.

Due to the fact that the data for completion of the CFAs had been collected previous to the discovery of the Monte Carlo simulation method of assessing adequate sample size, the simulation was completed after the analysis in order to assess the power of the analysis with the given sample size. It would arguably be better to complete the simulation prior to collecting data for the purposes of conducting a CFA in order to know whether it is possible to survey the appropriate number of participants. Difficulty was encountered in the Monte Carlo simulation process as neither Brown (2006) nor Muthén and Muthén (2002), the two sources that included examples of how to conduct a Monte Carlo simulation using the MPlus software for this purpose, included examples using ordinal data. It was therefore necessary to create syntax for use in MPlus which involved contacting the MPlus software support via both the online message board and email. The time and resources involved in the process of correctly completing the Monte Carlo simulation may prevent many evaluation professionals from attempting this portion of the analysis. If example syntax is available in the literature to match the type of data being used in the CFA, the process is not as lengthy since it simply involves changing the sample syntax.

However, if the data characteristics are unique enough that an example does not appear in the literature, then the Monte Carlo simulation process may add hours to the process, depending on how easily the syntax is adapted and the availability of support from the factor analysis software author. Example syntax for the simulation used in this study are included in Appendix C.

Guideline #1: If prior research (e.g. EFA or CFA results) exist for the scales being studied, and adequate time and resources are available, conduct a sample size analysis using Monte Carlo simulation to ascertain the sample size needed to obtain adequate power (.8). This analysis will provide evidence of sufficient power for the CFA. If this sample size is unobtainable, do not conduct a CFA.

2. Model complexity

Model complexity is an important consideration in CFAs with ordinal data because of its relationship to sample size and the subsequent impact that sample size has on the WLSMV parameter estimation method (Harrington, 2009). This fact provides another argument in favor of completing a Monte Carlo simulation to determine necessary sample size before obtaining the sample so that the researcher may know beforehand whether a CFA is advisable given the potential finite population size. The complexity of the model is part of the simulation, which makes this method much better than sample size rules of thumb, which were not determined for specific models.

The complexity of the models measured by the Evaluation Use and Evaluation Involvement Scales was greater than that of the majority of examples

presented in the CFA books by Brown (2006) and Harrington (2009). Understandably, due to editing constraints and ease of explanation, textbooks and guides on the topic will present examples with limited model complexity. However, one simulation study referenced for this research studied the impact of model complexity on CFA of categorical data, using a model with sixteen items and four factors as the most complex model (Potthast, 1993). In contrast, the most complex model in this research study was the model of the Evaluation Use Scales based on theory, which had twenty-five items and four factors. Perhaps the Evaluation Use Scales exceed the level of model complexity that should be assessed using CFA, but, as presented in the literature review, the concept is a complex one and therefore its measure should be as well. Henry and Mark (2003) provide a compelling argument for the fact that use is important in evaluation, but the field lacks validity evidence for measures of the concept, perhaps due in part to its complexity. Lacking a measure of the concept leaves the field of evaluation without sufficient research evidence about the impacts of use. Theories are important in guiding research, but they need to be tested with measurement tools of some sort. Without validity evidence to support these measurement tools, results of research conducted with them remains in question.

3. Categories per item

Generally a researcher has no control over the number of categories per item in the ordinal data being used for a CFA as this characteristic is usually determined when the measurement instrument is created. There has been some

research related to the effects on factor analysis of varying categories per item, with the suggestion that five or more are needed (Dolan, 1994). However, it appears to be common practice to complete CFA with ordinal data that have fewer categories. The impacts on the CFA results are not clearly understood.

When a researcher does have control over the number of categories per item used in the ordinal data, it may be advisable to have five or more categories if potential exists for future CFA of the data. However, this measurement decision should be made by with the bigger picture in mind. There may be other, more compelling reasons guiding the division of categories when designing the ordinal items that will measure the latent variable under consideration.

In the case of the Evaluation Use and Evaluation Involvement Scales, the survey tool was created in order to measure use and involvement in various aspects of large, multi-site evaluations of NSF-funded programs. At the time of the survey's creation, there were no plans to conduct CFAs on data collected with the instrument. Therefore no consideration was given to the number of categories per survey item in terms of the impacts on future CFAs conducted using the data. Had the team planned for CFAs of data collected with the Scales and known about the suggestion to use data with five or more categories per item in CFAs, it likely would have chosen to use five categories. However, the team did discuss the use of an even number of choices to avoid respondents' tendency to select the middle category, which has been suggested by survey development research (Nardi, 2006). Given these two considerations, then,

perhaps six categories would have been used. This lack of pre-planning for future CFAs is surely common in evaluation survey development. Decisions regarding the number of categories per item may impact the results of the survey as well as future factor analysis of the model for which the data were designed to measure.

Analysis Options

After exploration of a number of software packages for use in CFA, including MPlus, LISREL, and R, the MPlus software package was selected for this analysis due to its inclusion of the WLSMV parameter estimation method, which is recommended in the literature for use with ordinal data (Flora & Curran, 2004; Brown, 2006; Muthén & Muthén, 2007).

Guideline #2: Consider the characteristics of the data being used in the CFA and choose aspects of the analysis accordingly. For survey data using ordinal, intensity response format items, use the MPlus software (others may also work, but MPlus appears to have the best analysis options for this type of data).

1. Parameter estimation method

When a CFA is conducted, model parameters, or the characteristics of the population, are estimated and tested. Parameters include factor loadings, factor variance, and error variance. The maximum likelihood (ML) method of estimating these parameters is most often used when conducting a CFA (Harrington, 2009). "ML aims to find the parameter values that made the observed data most likely (or conversely, maximize the likelihood of the parameters given the data)"

(Brown, 2006, p. 73). In the MPlus factor analysis software, ML is the default option for estimation of parameters. However, ML is not appropriate to use with ordinal data (Harrington, 2009). For ordinal data, the robust weighted least squares (WLSMV) method of estimating parameters is recommended (Flora & Curran, 2004; Brown, 2006; Muthén & Muthén, 2007).

The weighted least squares estimation method is closely related to the generalized least squares (GLS) method, which "minimizes the discrepancy between the observed and predicted covariance matrices" (Brown, 2006, p. 387). In contrast to GLS, WLS uses a weight matrix based on estimates of the variances and covariances of each observed element (Brown, 2006). WLSMV is a specific type of WLS estimation that provides parameters "using a diagonal weight matrix (W) and robust standard errors and a mean- and variance-adjusted X^2 test statistic" (Brown, 2006, p. 388). Though WLS estimators require larger sample sizes than ML estimators do, WLSMV is not as sensitive to sample size as is WLS and due to its use of robust standard errors and adjusted chi square, it is the best option for CFAs with ordinal data.

The MPlus User's Guide (Muthén & Muthén, 1998-2007) provides guidance in conducting a CFA with ordinal data and Brown (2006) includes examples of analysis with this type of data. In contrast to the sample size simulation, the syntax for completion of the CFAs was relatively easy to write as the examples provided clear guidance and the error messages provided in the software lead the researcher toward necessary syntax changes.

Guideline #3: For survey data using ordinal, intensity response format items, use the robust weighted least squares (WLSMV) parameter estimation method.

2. Measures of correlation

One of the reasons the ML estimation method does not perform well with ordinal data, particularly ordinal data with five or fewer categories per item, is that it uses the product-moment correlation which works well with continuous, but not ordinal, data (Flora & Curran, 2004). Polychoric correlation, the correlation measure used with the WLSMV parameter estimation method in MPlus, "estimates the linear relationship between two unobserved continuous variables given only observed ordinal data" (Flora & Curran, 2004, p. 467). This measure, then, assumes the unobserved values are based on an underlying continuous distribution, but allows CFA to be performed on ordinal data. Other methods that do not utilize polychoric correlation require continuous data.

Guideline #4: For survey data using ordinal, intensity response format items, use the polychoric correlation measure, which is automatically paired with the WLSMV parameter estimation method in MPlus.

3. Measures of fit:

The results of CFA cannot determine whether the model being tested is the "correct" model for the given items and factors. It can only, through measures of fit, allow the researcher to determine whether the model exhibits "good fit" based on a predetermined set of suggested cutoff criteria. It is recommended that multiple measures of fit be considered in order to determine

whether good fit is achieved. Though the literature varies somewhat in recommendations of which fit measures to use in particular cases, and even what to call the categories of fit measures, the following list summarizes recommended fit measures to be used with ordinal data CFA in the categories of absolute fit, parsimony correction, and comparative fit.

- Absolute fit:

Chi square is the most commonly used absolute fit index, though it is sometimes referred to as a lack of fit index because good model fit is indicated by a non-significant ($p > .05$) result. As a measure, it "assesses the magnitude of discrepancy between the sample and fitted covariance matrices" (Hu & Bentler, 1999, p. 2). However, as outlined previously, chi square is sensitive to sample size in that, given a large enough sample, chi square will almost always be significant no matter the results of the CFA (Harrington, 2009). This fact was highlighted in the results of this study where the chi square values from the four CFAs ranged from 175.2 to 367.5 and were all significant ($p < .001$). Given the other evidence of lack of good fit, however, this result could be evidence of the issues with the chi square test or poor model fit in the four models assessed by the CFAs.

The root mean square residual (RMR) family of fit indices includes RMR, the standardized root mean square residual (SRMR) and the weighted root mean square resideual (WRMR). They are the "square root of the difference between the residuals of the sample covariance matrix and the hypothesized covariance

model" (Hooper, Coughlan, & Mullen, 2008, p. 54). WRMR uses the weighted difference. It is suitable for use with ordinal data and is the absolute fit index reported in MPlus when WLSMV estimation is used. The WRMR values for the CFAs in this study ranged from 1.3 to 1.6, all of which exceeded the recommended cutoff of 1.0 for evidence of good model fit. It is possible that the RMR or SRMR values, if available, would have met their respective cutoff values, but the MPlus software reports only WRMR with the WLSMV parameter estimation method, since its authors feel that it is the most reliable version of the RMR fit index to use with ordinal data (Muthén & Muthén, 1998-2007).

Guideline #5: For survey data using ordinal, intensity response format items, use the weighted root mean square residual (WRMR) as the absolute fit measure considered with a cutoff value of ≤ 1.0 (Yu, 2002),. The standardized root mean square residual (SRMR) was recommended in the literature (Hooper, Coughlan, & Mullen, 2008), but WRMR is reported in MPlus (Muthén & Muthén, 1998-2007).

- Parsimony correction:

Root mean square error of approximation (RMSEA) suggests "how well the model, with unknown but optimally chosen parameter estimates would fit the populations [sic] covariance matrix" (Hooper, Coughlan, & Mullen, 2008, p. 54). It is listed as a parsimony correction index by Brown (2006) due to the fact that it will select the model with the fewer number of parameters, thus favoring the parsimonious model. Hutchinson and Olmos (1998) recommend RMSEA

specifically for use with categorical data. The RMSEA values in this study ranged from .138 to .183, which were all above the recommended cutoff of .06 for evidence of goodness of fit. Though the literature warns that the index values are for comparison to the cutoff values and not to one another, it is interesting to note that, in the case of the Evaluation Involvement Scales models, the less parsimonious model, created based on theory, resulted in a smaller RMSEA value than did the more parsimonious model, which was based on the EFA results completed on data from the scales that suggested two fewer items and one fewer factor.

Guideline #6: For survey data using ordinal, intensity response format items, use the root mean square error of approximation (RMSEA) as the parsimony correction fit measure considered with a cutoff value of ≤ .06 (Hu & Bentler, 1999).

- Comparative fit:

Comparative fit indices compare the chi square value to a baseline model in which the null hypothesis that there is zero correlation between variables (Hooper, Coughlan, & Mullen, 2008). The comparative fit index is the most recent iteration of this set of indices and has been revised to perform well with small sample sizes, so it is the most widely used comparative fit index (Hooper, Coughlan, & Mullen, 2008). While the use of this index is not specific to ordinal data, its good performance with small sample sizes and wide use in CFAs in general recommends it for use in CFAs with ordinal data. This widely used fit

index was the only one for which any of the models showed evidence of good fit. The results ranged from .946 to .962, with three of the four models exceeding the recommended cutoff value of .95. However, meeting the cutoff for only one of three fit indices does not provide sufficient evidence of good model fit.

Guideline #7: For survey data using ordinal, intensity response format items, use the comparative fit index (CFI) as the comparative fit measure considered with a cutoff value of ≥ .95 (Hu & Bentler, 1999).

Other Guidelines

During the completion of a CFA, it is recommended that localized areas of ill fit be assessed in order to pinpoint areas in which the model may be adjusted in order to achieve better fit. However, these adjustments should not be made solely on the basis of CFA results, but rather because theory or prior research suggests them or can sufficiently explain why the adjustment may result in better fit (Brown, 2006). These adjustments can involve the removal of items or combinations of factors, but perhaps more frequently involve the loading of items on multiple factors or the free estimation of parameters. Details such as these are less likely to have been addressed in previous research, including EFAs conducted on the models. In the case of the Evaluation Use and Evaluation Involvement Scales, EFAs had indicated the removal of two items and one factor for each of the scales. The Beyond Evaluation Use grant team agreed with the indicated adjustments and these new models were assessed as part of the CFA

process detailed in this study. Possible free estimation of parameters or loading of items on multiple factors, which may result in better model fit, would perhaps be less likely to have shown up during the EFA process. When adjustments are made in the process of conducting a CFA, the CFA actually becomes an EFA as model changes are exploratory in nature. Finding better fit from this process necessitates the collection of another sample on which to conduct a CFA to provide validity evidence for this new model.

As previously suggested in describing the Monte Carlo sample size simulation guideline, conducting a CFA in the recommended manner may be prohibitive for evaluators. Given the fact that sample size may limit the ability to conduct the first CFA, it would necessarily follow that collecting a second sample (which could actually be a third sample if data were collected for an EFA) would make the process even more prohibitive. Time and resource constraints will also limit the possibility of conducting multiple CFAs. Funding for the evaluative process may run out well before the multiple factor analyses can be concluded. The Beyond Evaluation Use grant, for which the Evaluation Use and Evaluation Involvement Scales were created, was a multi-year research study that included a team of two faculty members and multiple graduate student research assistants. Even with multiple people working on the creation of the sets of scales, including the process of EFA and CFA of the data, the results took years to complete due to the length of time it takes to conduct online surveys and the resulting analyses of the data. Most evaluators do not have the luxury of hiring

graduate students or other assistants to conduct the analyses, nor are they able to wait years for the results.

Guideline #8: As with all CFAs, check localized areas of ill fit. Items of concern would be those with standardized residuals greater than or equal to the absolute value of 1.96 and modification indices greater than 4.0. Adjustments to the model should be based on more than just these indices and necessitate further CFAs to confirm the new model.

Given the lack of quantitative research in the areas of evaluation use and involvement in evaluations by somewhat unintended users, the only data-based information available to inform the CFAs conducted for this study were the results of the EFAs conducted on the Evaluation Use and Evaluation Involvement Scales. For this reason, the adjustments made to the Scales after the EFAs, in each case involving removal of two items and the decrease of factors by one, remained in question. It was therefore the plan in this research study to conduct CFAs on both the models based on theory and the models based on the results of the EFAs. If one or the other produced evidence of goodness of fit, it was thought that this model would be the better of the two. However goodness of fit indices cannot be used for comparison as they include only recommended cutoffs, beyond which a model is not considered to fit the data well. The two versions could not be measured in a nested design since they involved changes in both number of items and number of factors. The EFAs suggested changes to

the model which were seen as improvements to the models. Therefore it would have been advisable simply to conduct CFAs on the models based on the EFA results and not include the theoretical models in the CFA process.

Guideline #9: Do not attempt to ascertain which of multiple models is the "best" by using only goodness-of-fit indices as they are not designed to compare models, but rather indicate good fit for a model when it meets a cutoff criterion. When models can be compared using a nesting technique, then testing which has better fit is possible.

Limitations

The team that developed the Evaluation Use and Evaluation Involvement Scales completed extensive research in the process of writing the items and theorizing the factors involved in the measurement of the two constructs. EFAs were conducted on the two sets of scales, resulting in two of the models tested in this CFA study. This being the first attempt to complete a CFA on data generated by the scales, little guidance exists for revision of the models. This lack of research on which to draw necessarily limits the ability of the researcher to make educated revisions to the model. As outlined previously, the process of revision requires additional CFAs to confirm the adjusted models so time and resources limit the ability to complete the entire confirmatory process if the first CFA does not result in good model fit.

Another potential limitation that may have impacted the results of the CFA is that orthogonal rotation methods, specifically varimax rotation, were used in

both EFAs. According to Brown (2006), orthogonal rotation is prone to producing misleading solutions when there is expected intercorrelation of factors, as is the case with the Evaluation Use and Evaluation Involvement Scales. It is also the case that EFA solutions generated using oblique rotation methods are more likely to generalize to CFA.

Implications

In 1937, when factor methods were relatively new, Thurstone warned of their potential misuse. Originally designed for use with continuous data, these methods have undergone many changes since Thurstone's warning was issued. The more options there are, however, the more potential for misuse or misapplication of the methods. The guidelines for conducting a CFA with ordinal data outlined previously will assist evaluators in assessing whether CFA is possible given the characteristics of their survey data and the potential sample size available to them. If a CFA is advisable, the guidelines provide further direction for planning and conducting the analysis.

Because of its connection to other data characteristics, such as model complexity and number of intensity response format levels, adequate sample size is a particularly important consideration in the CFA of ordinal data. Because evaluators often have a finite population available to survey, it is advisable to complete a sample size simulation prior to conducting the survey for the purpose of conducting a CFA. Once the necessary sample size has been calculated, the evaluator should consider whether that size is achievable, given the potential

number of respondents and expected non-response to the survey. Advances in CFA methods are decreasing the sample sizes needed to complete CFA with adequate power, but it is still an important factor to consider. Depending on the goals of the CFA, other analysis options are available and may be possible with smaller sample sizes.

Another implication related to the finite populations often available to evaluators is the number of samples necessary to conduct a factor analysis process. The process often begins with exploration of the factors (EFA), followed by a CFA completed on a new sample to potentially confirm the factors found during exploration. If, as in the case of this study, fit indices do not provide evidence of good model fit, changes must be made to the model and a new sample must be gathered in order to confirm these changes with a second CFA. This process can potentially continue as changes are made to the model. With finite populations from which to draw samples, this process can prove difficult, if not impossible. This fact again highlights the need to assess the feasibility of the process with the given population and explore other options, such as scaling, which transforms ordinal data in order to use it in more complex statistical analyses so as not to break the analysis rules put forth by Stevens (1946), or principal components analysis, which can be used to reduce a large set of measures to a smaller one for the purposes of more manageable analysis (Brown, 2006). These techniques do not take the place of CFA, but rather offer

options, depending on the goals of the researcher/evaluator developing the survey.

A big issue in the field of evaluation is the lack of measurement instruments. Each evaluation case is unique, so it is difficult to create measurement tools that are useful in a variety of situations and settings. Time and resource constraints for evaluation professionals also limit the creation of instruments with adequate reliability and validity evidence. The amount time and resources that went into the creation of the Evaluation Use and Evaluation Involvement Scales, including the theoretical research, item writing, pilot testing, survey administration, EFAs, and CFAs is well beyond that available to most evaluators. Therefore most evaluation survey data has likely not undergone analyses to test the fit of the models on which the data is based. This fact leads one to question the validity of the results of these evaluations.

Even given a highly unusual amount of time and resources, the Beyond Evaluation Use team was unable to provide adequate validity evidence for the use of the Evaluation Use and Evaluation Involvement Scales. What do these results mean for the data previously collected and used in the EFAs and CFAs or for evaluations in general without the resources available to the Beyond Evaluation Use project? The results were summarized and conclusions about use and involvement were made, but perhaps they were made based on incorrect models. If these data should not be analyzed prior to the provision of validity evidence for the use of the Scales, then the analytical process will take

even longer than previously thought. Also, participants will be completing surveys solely for the purposes of conducting factor analyses, and not for the purposes of conducting an evaluation. This study highlights the fact that surveys measured with intensity response format items, considered ordinal data for analysis purposes, are perhaps even more questionable as nothing more than medians and percentiles can be summarized with their results until factor analysis or scaling is conducted to provide more options. Though the measurement literature provides guidance when it comes to the creation and use of surveys, practically speaking it can be difficult to follow when conducting evaluations.

Future Research

The field of evaluation lacks instruments to measure the use of, and involvement in, evaluations, as shown by Toal et. al. (2006) and Toal (2007). Following up on the EFAs previously conducted on data gathered with the Evaluation Use and Evaluation Involvement Scales, this set of CFAs attempted to provide further validity evidence in support of the factor structures proposed by the EFAs. Good model fit was not found in this study, pointing to the need for further revisions of the models. In order to better understand and model the factors underlying the concepts of evaluation use and evaluation involvement, revisions to the models found during the EFA process, guided by the results of the CFAs, particularly the modification indices, could be proposed and assessed. If evidence of good model fit is found with these revised models, data should be gathered from a new sample and a second set of CFAs should be conducted to

provide further validity evidence for the factor structure. The population could then be expanded to include individuals involved in other types of evaluations, beyond large, multi-site examples as were sampled for the EFAs and for this study.

Because the main purpose of this study was to better understand and articulate the data characteristics and analysis decisions necessary for completion of CFA with ordinal data, it is also important to consider the need for further research to guide this line of study. Specifically, more research is needed to understand the benefits of using the WRMR fit index as it is a fairly new addition to the list of fit indices used in CFA (Yu, 2002). More research is needed on the impacts to CFA when ordinal items are measured with four categories, as they were in the scales used in this study. Factor analysis techniques were designed for use with continuous data, and simulation studies designed to understand the outcome of analyses conducted with ordinal data must consider specific characteristics individually. Dolan (1994) cautioned against conducting CFAs on items with fewer than five categories, but it appears to be common practice to do so. Perhaps Thurstone (1937) portended this future in which factor analysis is continually misused, or perhaps the effects of using data with these characteristics are minimal. Further research is needed.

Given the amount of time and resources it took the Beyond Evaluation Use team of researchers and graduate students to complete the process of researching, creating, administering, summarizing, and factor analyzing the

Evaluation Use and Evaluation Involvement Scales, the ability of most evaluation settings to complete such processes is doubtful. The measurement literature reviewed for this study suggests that data gathered without completing this process may not show evidence of validity. Indeed, even when the process is completed there may be a lack of validity evidence, as in the case of this study where the CFAs did not result in good model fit. The feasibility of completing the factor analysis process on surveys created specifically for evaluations needs to be better understood as do the implications on evaluation results for not providing validity evidence for models on which evaluation findings are based.

REFERENCES

Alkin, M. C., Daillak, R., & White, P. (1979). *Using evaluations: Does evaluation make a difference?* Beverly Hills, CA: Sage.

Alkin, M. C., & Taut, S. M. (2003). Unbundling evaluation use. *Studies in Educational Evaluation,* 29, 1-12.

Allison, P. D. (2002). *Missing Data.* Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-136. Thousand Oaks, CA: Sage.

Babakus, E., Ferguson, C. E., & Jöreskog, K. G. (1987). The sensitivity of confirmatory maximum likelihood factor analysis to violations of measurement scale and distribution assumptions. *Journal of Marketing Research, 37,* 72-141.

Baker, B. O., Hardyck, C. D., & Petrinovich, L. F. (1966). Weak measurements vs. strong statistics: An empirical critique of S. S. Stevens' proscriptions on statistics. *Educational and Psychological Measurement*, 26, 291-309.

Bentler, P. M., (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107(2), 238-246

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

Burke, B. (1998). Evaluating for a change: Reflections on participatory methodology. *New Directions for Evaluation,* 80, 43-56.

Carifio, J., & Perla, R. J. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes. *Journal of Social Sciences*, 3, 106-116.

Cousins, J. B. (2003). Utilization effects of participatory evaluation. In T. Keleghan & D. L. Stufflebeam (Eds.), *International Handbook of Educational Evaluation*. Boston: Kluwer Academic Publishers, 245-266.

Cousins, J. B., & Whitmore, E. (1998). Framing participatory evaluation. In E. Whitmore (Ed.), *Understanding and practicing participatory evaluation* (pp. 5-24). New Directions for Evaluation (No. 80). San Francisco: Jossey-Bass.

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

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

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods,* 4, 466-491.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, 7, 286-299.

Fowler, F. J. (2009). *Survey research methods.* Thousand Oaks, CA: Sage Publications, Inc.

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.

Greene, J. (1988). Stakeholder participation and utilization in program evaluation. *Evaluation Review*, 12(2), 91-116.

Harrington, D. (2009). *Confirmatory factor analysis.* New York: Oxford University Press.

Harwell, M. R., & Gatti, G. G. (2001). Rescaling ordinal data to interval data in educational research. *Review of Educational Research,* 71(1), 105-131.

Henry, G. T., & Mark, M. M. (2003). Beyond use: Understanding evaluation's influence on attitudes and actions. *American Journal of Evaluation*. 24(3), 293-314.

Hooglund, J. J., & Boomsma, A. (1998). Robustness studies in covariance structural modeling: An overview and a meta-analysis. *Sociological Methods & Research,* 26, 329-367.

Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural equation modeling: Guidelines for determining model fit. *The Electronic Journal of Business Research Methods*, 6(1), 53-60.

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

Hutchinson, S. R., & Olmos, A. (1998). Behavior of descriptive fit indexes in confirmatory factor analysis using ordered categorical data. *Structural Equation Modeling,* 5, 344-364.

Igo, S. E. (2007). *The averaged American: Surveys, citizens, and the making of a mass public*. Cambridge, MA: Harvard University Press.

Jamieson, S. (2004). Likert scales: How to (ab)use them. *Medical Education*, 38, 1217-1218.

Johnson, K. (2008). *Developing and validating scales of evaluation use in multi-site STEM education evaluations.* American Evaluation Association Annual Meeting, Denver, Colorado.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34, 183-202.

Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research,* 36, 347-387.

King, J. A., Greenseid, L., Johnson, K., Lawrenz, F., Toal, S., & Volkov, B. (2007). *Initial results from "Beyond Evaluation Use": A study of involvement and influence in large, multi-site National Science Foundation (NSF) evaluations.* American Evaluation Association Conference, Baltimore, Maryland.

King, J. A., Lawrenz, F, Toal, S., Johnson, K., Roseland, D., & Johnson, G. (2009). *Making the Most of Multisite Evaluations.* American Evaluation Association/Centers for Disease Control and Prevention Summer Institute, Atlanta, Georgia.

Kirkhart, K. (2000). Reconceptualizing evaluation use: An integrated theory of influence. In V. Caracelli & H. Preskill (Eds.), *The expanding scope of evaluation use*. (pp. 5-23). New Directions for Evaluation (No. 94). San Francisco: Jossey-Bass.

Knapp, T. R. (1990). Treating ordinal scales as interval scales: An attempt to resolve the controversy. *Nursing Research*, 39, 121-123.

Kuh, G. D., & Umbach, P.D. (2004). College and character: Insights from the National Survey of Student Engagement. *New Directions for Institutional Research*, 122, 37-54.

Labovitz, S. (1967). Some observations on measurement and statistics. *Social Forces,* 46, 151-160.

Lawrenz, F., & King, J. A. (2009). Report 8: Beyond evaluation use cross case report.

Lawrenz, F., King, J. A., & Greenseid, L. O. (2005). *The relationship between involvement and use/influence in large multi-site evaluations.* American Evaluation Association and Canadian Evaluation Society Conference, Toronto, Ontario, Canada.

Lawrenz, F., King, J. A., Toal, S., Johnson, K., & Roseland, D. (2009). *Relationships between involvement and use in the context of multi-site evaluation.* American Evaluation Association Annual Conference, Orlando, Florida.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140, 1-55.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, 4(1), 84-99.

Marcus-Roberts, H. M., & Roberts, F. S. (1987). Meaningless statistics. *Journal of Educational Statistics,* 12, 383-394.

Mark, A. E., & Boruff-Jones, P. D. Information literacy and student engagement: What the National Survey of Student Engagement reveals about your campus. *College and Research Libraries*, 64, 480-493.

Muthén, B., & Kaplan D. (1985). A comparison of some methodologies for the factor-analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology,* 38, 171-180.

Muthén, B., & Kaplan D. (1992). A comparison of some methodologies for the factor analysis of non-normal Likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology,* 45, 19-30.

Muthén, L.K. and Muthén, B.O. (1998-2007). Mplus User's Guide. Fifth Edition. Los Angeles, CA: Muthén & Muthén

Muthén, L.K. and Muthén, B.O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, 9(4), 599-620

Nardi, P. M. (2006). *Doing survey research: A guide to quantitative methods.* (2nd ed.). Boston, MA: Pearson and Allyn and Bacon.

O'Brien, R. M. (1979). The use of Pearson's with ordinal data. *American Sociological Review,* 44, 851-857.

Patton, M. Q. (2008). *Utilization-focused evaluation* (4th ed.). Thousand Oaks, CA: Sage.

Potthast, M. J. (1993). Confirmatory factor analysis of ordered categorical variables with large models. *British Journal of Mathematical and Statistical Psychology*, 46, 273-286.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practice.* Newbury Park, CA: Sage Publications.

Sharma, S., Durvasula, S, & Dillon, W. R. (1989). Some results on the behavior of alternate covariance structure estimation procedures in the presence of non-normal data. *Journal of Marketing Research,* 26, 214-221.

Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology*, 15, 201-293.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103, 677-680.

Stevens, S. S. (1955). On the averaging of data. *Science,* 121, 113-116.

Thorndike, R. M. (2005). *Measurement and evaluation in psychology and education.* (7th ed.). Upper Saddle River, NJ: Pearson Education, Inc.

Thurstone, L. L. (1935). *The vectors of mind: Multiple-factor analysis for the isolation of primary traits.* Chicago: The University of Chicago Press.

Thurstone, L. L. (1937). Current misuse of the factorial methods. *Psychometrika*, 2, 73-76.

Toal, S. A. (2007). The development and validation of an evaluation involvement scale for use in multi-site evaluations. Unpublished doctoral dissertation, University of Minnesota, Twin Cities.

Toal, S. A., Greenseid, L., Lawrenz, F., King, J. A., & Johnson, K. (2006). *Researching evaluation use and influence: Twenty years of empirical study*. American Evaluation Association Annual Conference, Portland, Oregon.

Wang, L., Fan, X., & Willson, V. L. (1996). Effects of non-normal data on parameter estimates and fit indices for a model with latent and manifest variables: An empirical study. *Structural Equation Modeling,* 3*,* 228-247.

Weiss, C. H. (1998). *Evaluation research: Methods for studying program and policies* (2nd ed.). Upper Saddle River, NJ: Prentice-Hall Incorporated.

Yu, C. Y. (2002). Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes. Unpublished doctoral dissertation, University of California, Los Angeles.

APPENDIX A - Survey Emails

Pre-notice email

Dear Dr. (*Insert Last Name*),

I am Gina Johnson, a doctoral student in Educational Psychology at the University of Minnesota. As part of a research project examining the use and influence of NSF program evaluations, I am asking for your help.

The National Science Foundation (NSF) seeks to improve the use of the program evaluations completed on its grants. As a recent recipient of a (*Insert name of program*) grant, you are uniquely qualified to provide essential and firsthand input about your experiences with the (*Insert name of program*) program evaluation.

In about one week you will receive an email containing directions on how to participate in an online survey designed to help us better understand the NSF evaluation process and its outcomes. The survey will take approximately 10 minutes to complete. Your responses to the survey will also enable NSF to enhance its evaluation activities.

We thank you in advance for your willingness to help with this project. Your input is critical, and we appreciate your involvement.

Sincerely,

Gina Johnson
Doctoral Student, Educational Psychology – Quantitative Methods in Education
University of Minnesota
john3673@umn.edu

This study is sponsored by NSF number REC0438545
IRB# 0709E16085

Invitation to participate in survey email

Dear Dr. (*Insert Last Name*),

I am Gina Johnson, a doctoral student in Educational Psychology at the University of Minnesota. As part of a research project examining the use and influence of NSF program evaluations, I would like to ask for your help.

The National Science Foundation (NSF) seeks to improve the use of the program evaluations completed on its grants. As a recent recipient of an (*Insert name of program*) grant, you are uniquely qualified to provide essential and firsthand input about your experiences with the (*Insert name of program*) program evaluation. The linked survey requests candid perceptions of your experience, which will help us better understand the evaluation process and its outcomes. This information will enable NSF to enhance its evaluation activities.

No names will be used in this study, and only aggregated results will be reported. At the end of the survey, you will be asked for permission to contact you if you are selected for a follow-up interview. The web-based survey takes approximately 10 minutes to complete.

To participate, click on the survey link below, and you will be prompted for your pre-assigned account name and password (provided at the end of this email).

Please feel free to contact me at the email address listed below if you have questions or concerns about this survey. Or, if you would like to speak to someone else, you may contact the University of Minnesota's Research Subjects' Advocate Line at 612-625-1650.

We recognize the many demands made on your time, and thank you in advance for your willingness to help. Your input is critical, and we appreciate your involvement.

Sincerely,

Gina Johnson

Doctoral Student, Educational Psychology – Quantitative Methods in Education
University of Minnesota
john3673@umn.edu

Link:
Your account name:

Your password:

This study is sponsored by NSF number REC0438545
IRB# 0709E16085

Reminder email

Dear Dr. (*insert last name*),

You have recently received an emailed invitation to complete a survey about the use and influence of NSF program evaluations. This survey will close on May 26th, 2009.

In order to understand the evaluation process, it is important that we hear from persons with all levels of involvement in the national evaluation of your program, both PIs and co-PIs. Each person's perspective is important and we hope to hear from all of you.

To participate, please click on the survey link below, and you will be prompted to enter the following account name and password:

Link:
Your account name:
Your password:

If you have any questions or concerns, please feel free to contact us at the email address listed. Or, if you would like to talk to someone else, you may contact the University of Minnesota's Research Subjects' Advocate Line at (612) 625-1650.

We recognize the many demands made on your time, and greatly value your participation in the project.

Sincerely,

Gina Johnson
Doctoral Student, Educational Psychology – Quantitative Methods in Education
University of Minnesota
eug@umn.edu

This study is sponsored by NSF number REC0438545
IRB# 0709E16085

Non-response survey email

Subject Line: Survey Nonresponse Assistance Request

A few months ago, we emailed you a survey about your participation with evaluation activities related to your work with (*Insert name of program*), but we did not receive a response from you. That survey is closed. Now, we are asking for your help to learn why some people responded and others did not. Your input is critical to our assessment of possible non-respondent bias.

Just hit reply to respond to the three quick questions below. Put an "X" next your answer choice for each question.

Thank you for your help.

1. How involved were you in the (*Insert name of program*) evaluation activities?

___ Not at all involved.
___ Involved a little.
___ Involved some.
___ Involved extensively.

2. How much impact did the (*Insert name of program*) evaluation activities have on you?

___ No impact.
___ A little impact.
___ Some impact.
___ Extensive impact.

3. What was the main thing that kept you from responding to our initial survey request? (Pick one.)

___Not enough time.
___Do not remember receiving it.
___Did not feel it applied to me.
___Technical problems.
___Other (please specify):

Thank you again for taking the time to help ensure the quality of this study.

This study by the University of Minnesota is sponsored by NSF number REC0438545
IRB# 0709E16085

APPENDIX B – MPlus CFA Syntax

Model: Use from Theory

```
TITLE:      INFLUENCE AND USE CFA from Theory

DATA:       FILE IS
C:\Documents and Settings\Sean\Desktop\MPLUS\Influence Use_Theory.dat;

VARIABLE:   NAMES ARE
                Account q025 q026 q027 q028 q029 q030 q031 q032 q033 q034 q035
                q036 q038 q039 q040 q041 q042 q043
                q050 q051 q052 q053 q054 q055 q056 q057 q058;
            CATEGORICAL ARE
                q025 q026 q027 q028 q029 q030 q031 q032 q033 q034 q035
                q 036 q038 q039 q040 q041 q042 q043
                q050 q051 q052 q053 q054 q055 q056 q057 q058;
            USEVARIABLES ARE
                q025 q026 q027 q028 q029 q030 q031 q032 q033 q034 q035
                q 036 q038 q039 q040 q041 q042 q043
                q050 q051 q052 q053 q054 q055 q056 q057 q058;
            MISSING ARE ALL (9);

ANALYSIS:   ESTIMATOR=WLSMV;
            ITERATIONS = 5000;

MODEL:      f1 BY q025 q026 q027 q028 q029 q030
            f2 BYq031 q032 q033 q034 q035 q036;
            f3 BY q038 q039 q040 q041 q042 q043;
            f4 BY q050 q051 q052 q053 q054 q055 q056 q057 q058;
            f1 WITH f2;
            f1 WITH f3;
            f1 WITH f4;
            f2 WITH f3;
            f2 WITH f4;
            f3 WITH f4;


OUTPUT:     MODINDICES(4.00) RESIDUAL STAND;
            PATTERNS;
```

Model: Use from EFA

```
TITLE:      INFLUENCE AND USE CFA from EFA

DATA:       FILE IS
C:\Documents and Settings\Sean\Desktop\MPLUS\Influence Use_EFA.dat;

VARIABLE:   NAMES ARE
                Account q025 q026 q027 q028 q029 q030 q031 q032 q033 q034 q035
                q038 q039 q040 q042 q043
                q050 q051 q052 q053 q054 q055 q056 q057 q058;
            CATEGORICAL ARE
                q025 q026 q027 q028 q029 q030 q031 q032 q033 q034 q035
                q038 q039 q040 q042 q043
                q050 q051 q052 q053 q054 q055 q056 q057 q058;
            USEVARIABLES ARE
                q025 q026 q027 q028 q029 q030 q031 q032 q033 q034 q035
                q038 q039 q040 q042 q043
                q050 q051 q052 q053 q054 q055 q056 q057 q058;
            MISSING ARE ALL (9);

ANALYSIS:   ESTIMATOR=WLSMV;
            ITERATIONS = 5000;

MODEL:      f1 BY q025 q026 q027 q028 q029 q030 q031 q032 q033 q034 q035;
            f2 BY q038 q039 q040 q042 q043;
            f3 BY q050 q051 q052 q053 q054 q055 q056 q057 q058;
            f1 with f2;
            f1 with f3;
            f2 with f3;


OUTPUT:     MODINDICES(4.00) RESIDUAL STAND;
            PATTERNS;
```

Model: Involvement from Theory

```
TITLE:      INVOLVEMENT CFA From Theory

DATA:       FILE IS
C:\Documents and Settings\Sean\Desktop\MPLUS\Involvement_Theory.dat;

VARIABLE:   NAMES ARE
                Account q001 q002 q003 q005 q006 q007 q008 q009 q010 q012 q 013
                q014 q015;
                CATEGORICAL ARE
                q001 q002 q003 q005 q006 q007 q008 q009 q010 q012 q013 q014
q015;
            USEVARIABLES ARE
                q001 q002 q003 q005 q006 q007 q008 q009 q010 q012 q013 q014
q015;
            MISSING ARE ALL (9);

ANALYSIS:   ESTIMATOR=WLSMV;
            ITERATIONS = 5000;

MODEL:      f1 BY q001 q002 q003;
            f2 BY q005 q006 q007 q008 q009 q010;
            f3 BY q012 q013 q014 q015;
            f1 with f2;
            f1 with f3;
            f2 with f3;

OUTPUT:     MODINDICES(4.00) RESIDUAL STAND;
            PATTERNS;
```

Model: Involvement from EFA

```
TITLE:      INVOLVEMENT CFA From EFA

DATA:       FILE IS
C:\Documents and Settings\Sean\Desktop\MPLUS\Involvement_EFA.dat;

VARIABLE:   NAMES ARE
                Account q001 q002 q003 q005 q006 q007 q008 q009 q010 q012 q014;
                CATEGORICAL ARE
                q001 q002 q003 q005 q006 q007 q008 q009 q010 q012 q014;
            USEVARIABLES ARE
                q001 q002 q003 q005 q006 q007 q008 q009 q010 q012 q014;
            MISSING ARE ALL (9);

ANALYSIS:   ESTIMATOR=WLSMV;
            ITERATIONS = 5000;

MODEL:      f1 BY q001 q002 q003 q005 q006;
            f2 BY q007 q008 q009 q010 q012 q014;
            f1 with f2;

OUTPUT:     MODINDICES(4.00) RESIDUAL STAND;
            PATTERNS;
```

APPENDIX C – MPlus Monte Carlo Sample Size Simulation Syntax

Model: Use from Theory

```
TITLE:          MC sample size simulation for Use Scale Theory
MONTECARLO:
                NAMES ARE X1-X27;
                GENERATE = X1-X27 (3 p);
                CATEGORICAL ARE X1-X27;
                NOBSERVATIONS = 296;
                NREPS = 10000;
                SEED = 333

MODEL POPULATION:

                KNOW BY X1*.956  X2*.955  X3*.946  X4*.888
                X5*.920  X6*.866;
                SKILL BY X7*.965  X8*.959  X9*.942  X10*.873
                X11*.925 X12*.925;
                BELIEF BY X13*.953  X14*.963  X15*.966  X16*.876
                X17*.939  X18*.861;
                FIND BY X19*.893  X20*.953  X21*.896  X22*.705
                X23*.855  X24*.864  X25*.800  X26*.894  X27*.892;

                KNOW@1; SKILL@1; BELIEF@1; FIND@1;
                X1*.086;  X2*.088;  X3*.105;  X4*.211;  X5*.154;
                X6*.250;  X7*.069;  X8*.080;  X9*.113;  X10*.238;
                X11*.144;  X12*.144; X13*.092;  X14*.073;
                X15*.067;  X16*.233;  X17*.118;  X18*.259;
                X19*.203;  X20*.092;  X21*.197;  X22*.503;
                X23*.269;  X24*.254; X25*.36;  X26*.201;  X27*.204;
                KNOW WITH SKILL*.918;
                KNOW WITH BELIEF*.643;
                KNOW WITH FIND*.576;
                SKILL WITH BELIEF*.661;
                SKILL WITH FIND*.543;
                BELIEF WITH FIND*.498;
```

```
[X1$1*-.822 X1$2*-.112 X1$3*.991
X2$1*-.832 X2$2*-.082 X2$3*1.061
X3$1*-.713 X3$2*-.052 X3$3*1.074
X4$1*-.463 X4$2*.241 X4$3*1.241
X5$1*-.761 X5$2*.022 X5$3*1.157
X6$1*-.931 X6$2*-.316 X6$3*.637
X7$1*-.680 X7$2*-.087 X7$3*1.029
X8$1*-.640 X8$2*.022 X8$3*1.108
X9$1*-.598 X9$2*.108 X9$3*1.174
X10$1*-.403 X10$2*.293 X10$3*1.264
X11$1*-.613 X11$2*.170 X11$3*1.241
X12$1*-.782 X12$2*-.026 X12$3*.945
X13$1*-.784 X13$2*-.221 X13$3*.619
X14$1*-.782 X14$2*-.139 X14$3*.658
X15$1*-.708 X15$2*-.157 X15$3*.742
X16$1*-.316 X16$2*.262 X16$3*.986
X17$1*-.551 X17$2*.039 X17$3*.890
X18$1*-.688 X18$2*-.100 X18$3*.756
X19$1*-.546 X19$2*-.074 X19$3*.683
X20$1*-.852 X20$2*-.322 X20$3*.531
X21$1*-.621 X21$2*-.113 X21$3*.765
X22$1*-.479 X22$2*-.105 X22$3*.610
X23$1*-.374 X23$2*.070 X23$3*.887
X24$1*-.282 X24$2*.078 X24$3*.824
X25$1*-.039 X25$2*.315 X25$3*1.037
X26$1*-.669 X26$2*-.130 X26$3*.817
X27$1*-.412 X27$2*-.026 X27$3*.824];

MODEL:
KNOW BY X1*.956  X2*.955  X3*.946  X4*.888
X5*.920  X6*.866;
SKILL BY X7*.965  X8*.959  X9*.942  X10*.873
X11*.925 X12*.925;
BELIEF BY X13*.953  X14*.963  X15*.966  X16*.876
X17*.939  X18*.861;
FIND BY X19*.893  X20*.953  X21*.896  X22*.705
X23*.855  X24*.864  X25*.800  X26*.894  X27*.892;
KNOW@1; SKILL@1; BELIEF@1; FIND@1;
KNOW WITH SKILL*.918;
KNOW WITH BELIEF*.643;
KNOW WITH FIND*.576;
SKILL WITH BELIEF*.661;
```

```
SKILL WITH FIND*.543;
BELIEF WITH FIND*.498;


        [X1$1*-.822 X1$2*-.112 X1$3*.991
        X2$1*-.832 X2$2*-.082 X2$3*1.061
        X3$1*-.713 X3$2*-.052 X3$3*1.074
        X4$1*-.463 X4$2*.241 X4$3*1.241
        X5$1*-.761 X5$2*.022 X5$3*1.157
        X6$1*-.931 X6$2*-.316 X6$3*.637
        X7$1*-.680 X7$2*-.087 X7$3*1.029
        X8$1*-.640 X8$2*.022 X8$3*1.108
        X9$1*-.598 X9$2*.108 X9$3*1.174
        X10$1*-.403 X10$2*.293 X10$3*1.264
        X11$1*-.613 X11$2*.170 X11$3*1.241
        X12$1*-.782 X12$2*-.026 X12$3*.945
        X13$1*-.784 X13$2*-.221 X13$3*.619
        X14$1*-.782 X14$2*-.139 X14$3*.658
        X15$1*-.708 X15$2*-.157 X15$3*.742
        X16$1*-.316 X16$2*.262 X16$3*.986
        X17$1*-.551 X17$2*.039 X17$3*.890
        X18$1*-.688 X18$2*-.100 X18$3*.756
        X19$1*-.546 X19$2*-.074 X19$3*.683
        X20$1*-.852 X20$2*-.322 X20$3*.531
        X21$1*-.621 X21$2*-.113 X21$3*.765
        X22$1*-.479 X22$2*-.105 X22$3*.610
        X23$1*-.374 X23$2*.070 X23$3*.887
        X24$1*-.282 X24$2*.078 X24$3*.824
        X25$1*-.039 X25$2*.315 X25$3*1.037
        X26$1*-.669 X26$2*-.130 X26$3*.817
        X27$1*-.412 X27$2*-.026 X27$3*.824];


ANALYSIS:
        PARAMETERIZATION = DELTA;
        ESTIMATOR = WLSMV;
OUTPUT:
        TECH9;
```

Model: Use from EFA

```
TITLE:          MC sample size simulation for Use Scale EFA
MONTECARLO:
                NAMES ARE X1-X25;
                GENERATE = X1-X25 (3 p);
                CATEGORICAL ARE X1-X25;
                NOBSERVATIONS = 296;
                NREPS = 10000;
                SEED = 3333

MODEL POPULATION:
                SKILL BY X1*.953  X2*.949  X3*.939  X4*.878  X5*.911
                X6*.839  X7*.961  X8*.955  X9*.935  X10*.854  X11*.916;
                BELIEF BY X12*.955  X13*.965  X14*.968  X15*.923
                X16*.844;
                FIND BY X17*.897  X18*.950  X19*.896  X20*.705  X21*.851
                X22*.866  X23*.799  X24*.894  X25*.893;
                SKILL@1; BELIEF@1; FIND@1;
                X1*.092;  X2*.099;  X3*.118;  X4*.229;  X5*.170;  X6*.296;
                X7*.076;  X8*.088;  X9*.126;  X10*.271;  X11*.161;
                X12*.088;  X13*.069;  X14*.063;  X15*.148;  X16*.288;
                X17*.195;  X18*.098;  X19*.197;  X20*.503;  X21*.276;
                X22*.250;  X23*.362; X24*.201;  X25*.203;
                SKILL WITH BELIEF*.671; SKILL WITH FIND*.558;
                BELIEF WITH FIND*.498;

                    [X1$1*-.822 X1$2*-.112 X1$3*.991
                    X2$1*-.832 X2$2*-.082 X2$3*1.061
                    X3$1*-.713 X3$2*-.052 X3$3*1.074
                    X4$1*-.463 X4$2*.241 X4$3*1.241
                    X5$1*-.761 X5$2*.022 X5$3*1.157
                    X6$1*-.931 X6$2*-.316 X6$3*.637
                    X7$1*-.680 X7$2*-.087 X7$3*1.029
                    X8$1*-.640 X8$2*.022 X8$3*1.108
                    X9$1*-.598 X9$2*.108 X9$3*1.174
                    X10$1*-.403 X10$2*.293 X10$3*1.264
                    X11$1*-.613 X11$2*.170 X11$3*1.241
                    X12$1*-.784 X12$2*-.221 X12$3*.619
                    X13$1*-.782 X13$2*-.139 X13$3*.658
                    X14$1*-.708 X14$2*-.157 X14$3*.742
                    X15$1*-.551 X15$2*.039 X15$3*.890
                    X16$1*-.688 X16$2*-.100 X16$3*.756
                    X17$1*-.546 X17$2*-.074 X17$3*.683
```

X18$1*-.852 X18$2*-.322 X18$3*.531
X19$1*-.621 X19$2*-.113 X19$3*.765
X20$1*-.479 X20$2*-.105 X20$3*.610
X21$1*-.374 X21$2*.070 X21$3*.887
X22$1*-.282 X22$2*.078 X22$3*.824
X23$1*-.039 X23$2*.315 X23$3*1.037
X24$1*-.669 X24$2*-.130 X24$3*.817
X25$1*-.412 X25$2*-.026 X25$3*.824];

MODEL:

SKILL BY X1*.953 X2*.949 X3*.939 X4*.878 X5*.911
X6*.839 X7*.961 X8*.955 X9*.935 X10*.854 X11*.916;
BELIEF BY X12*.955 X13*.965 X14*.968 X15*.923
X16*.844;
FIND BY X17*.897 X18*.950 X19*.896 X20*.705
X21*.851 X22*.866 X23*.799 X24*.894 X25*.893;
SKILL@1; BELIEF@1; FIND@1;
SKILL WITH BELIEF*.671; SKILL WITH FIND*.558;
BELIEF WITH FIND*.498;

[X1$1*-.822 X1$2*-.112 X1$3*.991
X2$1*-.832 X2$2*-.082 X2$3*1.061
X3$1*-.713 X3$2*-.052 X3$3*1.074
X4$1*-.463 X4$2*.241 X4$3*1.241
X5$1*-.761 X5$2*.022 X5$3*1.157
X6$1*-.931 X6$2*-.316 X6$3*.637
X7$1*-.680 X7$2*-.087 X7$3*1.029
X8$1*-.640 X8$2*.022 X8$3*1.108
X9$1*-.598 X9$2*.108 X9$3*1.174
X10$1*-.403 X10$2*.293 X10$3*1.264
X11$1*-.613 X11$2*.170 X11$3*1.241
X12$1*-.784 X12$2*-.221 X12$3*.619
X13$1*-.782 X13$2*-.139 X13$3*.658
X14$1*-.708 X14$2*-.157 X14$3*.742
X15$1*-.551 X15$2*.039 X15$3*.890
X16$1*-.688 X16$2*-.100 X16$3*.756
X17$1*-.546 X17$2*-.074 X17$3*.683
X18$1*-.852 X18$2*-.322 X18$3*.531
X19$1*-.621 X19$2*-.113 X19$3*.765
X20$1*-.479 X20$2*-.105 X20$3*.610
X21$1*-.374 X21$2*.070 X21$3*.887
X22$1*-.282 X22$2*.078 X22$3*.824
X23$1*-.039 X23$2*.315 X23$3*1.037

```
                X24$1*-.669 X24$2*-.130 X24$3*.817
                X25$1*-.412 X25$2*-.026 X25$3*.824];

ANALYSIS:
                PARAMETERIZATION = DELTA;
                ESTIMATOR = WLSMV;
OUTPUT:
                TECH9;
```

Model: Involvement from Theory

```
TITLE:          MC sample size simulation for Involvement Scale Theory
MONTECARLO:
                NAMES ARE X1-X13;
                GENERATE = X1-X13 (3 p);
                CATEGORICAL ARE X1-X13;
                NOBSERVATIONS = 296;
                NREPS = 10000;
                SEED = 3

MODEL POPULATION:
                PLAN BY X1*.942  X2*.944  X3*.957;
                IMPLEM BY X4*.913  X5*.957 X6*.720  X7*.846
                X8*.938  X9*.923;
                COMM BY X10*.860  X11*.905  X12*.892 X13*.893;
                PLAN@1; IMPLEM@1; COMM@1;
                X1*.113;  X2*.109;  X3*.084;  X4*.166;  X5*.084;  X6*.482;
                X7*.284;  X8*.120;  X9*.148;
                X10*.260;  X11*.181;  X12*.204;  X13*.203;
                PLAN WITH IMPLEM*.64;
                PLAN WITH COMM*.762;
                IMPLEM WITH COMM*.845;

                        [X1$1*-.844 X1$2*-.421 X1$3*.178
                        X2$1*-.327 X2$2*-.075 X2$3*.383
                        X3$1*-.466 X3$2*-.118 X3$3*.515
                        X4$1*-.135 X4$2*.286 X4$3*.852
                        X5$1*-.264 X5$2*.237 X5$3*.788
                        X6$1*-.264 X6$2*.184 X6$3*.621
                        X7$1*-.157 X7$2*.157 X7$3*.664
                        X8$1*-.044 X8$2*.275 X8$3*.882
                        X9$1*-.330 X9$2*.026 X9$3*.691
                        X10$1*.192 X10$2*.544 X10$3*1.043
                        X11$1*-.447 X11$2*.031 X11$3*.569
                        X12$1*-.130 X12$2*.112 X12$3*.569
                        X13$1*-.548 X13$2*-.229 X13$3*.220];

MODEL:
                PLAN BY X1*.942  X2*.944  X3*.957;
                IMPLEM BY X4*.913  X5*.957 X6*.720  X7*.846
                X8*.938  X9*.923;
```

```
COMM BY X10*.860  X11*.905  X12*.892 X13*.893;
PLAN@1; IMPLEM@1; COMM@1;
PLAN WITH IMPLEM*.64;
PLAN WITH COMM*.762;
IMPLEM WITH COMM*.845;

    [X1$1*-.844 X1$2*-.421 X1$3*.178
    X2$1*-.327 X2$2*-.075 X2$3*.383
    X3$1*-.466 X3$2*-.118 X3$3*.515
    X4$1*-.135 X4$2*.286 X4$3*.852
    X5$1*-.264 X5$2*.237 X5$3*.788
    X6$1*-.264 X6$2*.184 X6$3*.621
    X7$1*-.157 X7$2*.157 X7$3*.664
    X8$1*-.044 X8$2*.275 X8$3*.882
    X9$1*-.330 X9$2*.026 X9$3*.691
    X10$1*.192 X10$2*.544 X10$3*1.043
    X11$1*-.447 X11$2*.031 X11$3*.569
    X12$1*-.130 X12$2*.112 X12$3*.569
    X13$1*-.548 X13$2*-.229 X13$3*.220];

ANALYSIS:
    PARAMETERIZATION = DELTA;
    ESTIMATOR = WLSMV;
OUTPUT:
    TECH9;
```

Model: Involvement from EFA

```
TITLE:          MC sample size simulation for Involvement Scale EFA
MONTECARLO:
                NAMES ARE X1-X11;
                GENERATE = X1-X11 (3 p);
                CATEGORICAL ARE X1-X11;
                NOBSERVATIONS = 296;
                NREPS = 10000;
                SEED = 33;
MODEL POPULATION:
                PLAN BY X1*.924  X2*.922  X3*.923  X4*.914  X5*.962;
                IMPLEM BY X6*.741  X7*.853  X8*.956  X9*.923  X10*.855
                X11*.872;
                PLAN@1; IMPLEM@1;
                X1*.146;  X2*.150;  X3*.148;  X4*.164;  X5*.075;
                X6*.451;  X7*.272;  X8*.086;  X9*.148;  X10*.269;  X11*.240;
                PLAN WITH IMPLEM*.726;

                        [X1$1*-.844 X1$2*-.421 X1$3*.178
                        X2$1*-.327 X2$2*-.075 X2$3*.383
                        X3$1*-.466 X3$2*-.118 X3$3*.515
                        X4$1*-.135 X4$2*.286 X4$3*.852
                        X5$1*-.264 X5$2*.237 X5$3*.788
                        X6$1*-.264 X6$2*.184 X6$3*.621
                        X7$1*-.157 X7$2*.157 X7$3*.664
                        X8$1*-.044 X8$2*.275 X8$3*.882
                        X9$1*-.330 X9$2*.026 X9$3*.691
                        X10$1*.192 X10$2*.544 X10$3*1.043
                        X11$1*-.130 X11$2*.112 X11$3*.569];


MODEL:
                PLAN BY X1*.924  X2*.922  X3*.923  X4*.914  X5*.962;
                IMPLEM BY X6*.741  X7*.853  X8*.956  X9*.923  X10*.855
                X11*.872;
                PLAN@1; IMPLEM@1;
                PLAN WITH IMPLEM*.726;

                        [X1$1*-.844 X1$2*-.421 X1$3*.178
                        X2$1*-.327 X2$2*-.075 X2$3*.383
                        X3$1*-.466 X3$2*-.118 X3$3*.515
```

```
                X4$1*-.135 X4$2*.286 X4$3*.852
                X5$1*-.264 X5$2*.237 X5$3*.788
                X6$1*-.264 X6$2*.184 X6$3*.621
                X7$1*-.157 X7$2*.157 X7$3*.664
                X8$1*-.044 X8$2*.275 X8$3*.882
                X9$1*-.330 X9$2*.026 X9$3*.691
                X10$1*.192 X10$2*.544 X10$3*1.043
                X11$1*-.130 X11$2*.112 X11$3*.569];

ANALYSIS:       PARAMETERIZATION = DELTA;
                ESTIMATOR = WLSMV;
OUTPUT:
                TECH9;
```