# AN APPLICATION OF UNFOLDING AND CUMULATIVE ITEM RESPONSE THEORY MODELS FOR NONCOGNITIVE SCALING: EXAMINING THE ASSUMPTIONS AND APPLICABILITY OF THE GENERALIZED GRADED UNFOLDING MODEL

Adrienne N. Sgammato

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the School of Education.

Chapel Hill
2009

Approved by

Advisor: Dr. Gregory J. Cizek

Reader: Terry A. Ackerman

Reader: Gary T. Henry

Reader: Richard M. Luecht

Reader: William B. Ware

# ABSTRACT

Adrienne N. Sgammato: An Application of Unfolding and Cumulative Item Response Theory Models for Non-Cognitive Scaling: Examining the Assumptions and Applicability of the Generalized Graded Unfolding Model

(Under the direction of Gregory J. Cizek)

This study examined the applicability of a relatively new unidimensional, unfolding item response theory (IRT) model called the generalized graded unfolding model (GGUM; Roberts, Donoghue, & Laughlin, 2000). A total of four scaling methods were applied. Two commonly used cumulative IRT models for polytomous data, the Partial Credit Model and the Generalized Partial Credit model were considered. The third scaling approach was the use of a confirmatory factor analysis. The fourth model, an unfolding IRT model, the Generalized Graded Unfolding Model was considered. These models were applied to attitudinal data from 65,031 licensed teachers in North Carolina who responded to a survey about their working conditions. Two subscales (Empowerment and Leadership) were used and analyzed separately. Items are Likert-type with five response options ranging from Strongly Agree to Strongly Disagree.

Analyses focused on examination of the correspondence between the assumptions that underlie the data and the IRT models, revealing evidence about the structure of the data, the location of people and items, and the response process governing observed data. The analyses included graphical representations of person and item estimates as well as analytical examination of item characteristic curves (ICCs) for the various models.

Various indices of relative and absolute model fit statistics are presented for the IRT models. Although the two scales were originally built using factor analytic methods, results suggested that a single factor model did not fit the Empowerment well, though fit of the Leadership data was moderate. Tests of IRT model assumptions indicated that cumulative assumptions were meet more often than those that underlie unfolding IRT models. Comparison of item and person parameter estimates show that, across both scales, cumulative and unfolding IRT models functioned very similarly. However, some item on both scales did exhibit unfolding properties. Finally, a summary of potential extensions of the GGUM model and other contributions of this research including the possibility of using unfolding models for scale development and attitude measurement in areas beyond that of working conditions of teachers or administrators are offered.

# ACKNOWLEDGEMENTS

My accomplishments and success would not have been possible without the constant support and encouragement of my parents Ralph and Christine Sgammato, and my siblings Rafel and Joseph. They always believed in me. I certainly have had help along the way from family, friends, and colleagues. I am sincerely grateful for their endorsement and faith.

To my advisor, Dr. Gregory Cizek, I am especially appreciative of his support and guidance throughout my graduate career, and especially throughout the dissertation process. I am also indebted to my dissertation committee members, Dr. William Ware, Dr. Gary Henry, Dr. Richard Luecht, and Dr. Terry Ackerman for their insight, enthusiasm, and patience. The provision of technical expertise by Dr. Richard Luecht, Dr. James Roberts, Dr. Chris Weisen, and the statisticians at Scientific Software Inc. has been immensely helpful and insightful.

## TABLE OF CONTENTS

APPENDICES

# LIST OF TABLES

Table Page

# LIST OF FIGURES

Figure Page

# CHAPTER 1

# INTRODUCTION

Classical test theory (CTT) approaches to test development and scoring were predominantly used as early as the start of the 1900s within the context of intelligence testing. Contributions by Lawley (1943, 1944), Birnbaum (1957, 1958), Rasch (1960), Wright (1967), and Lord and Novick (1968) (as cited in Hambleton & Swaminathan, 1985) to psychometric/test theory evolved into what is known as item response theory (IRT). IRT measurement models are considered to be more sophisticated than CTT models in that IRT analyses allow for: 1) the estimation of performance on each item of a test for any examinee who is at a particular point on the underlying trait; 2) more precise measures of accuracy by estimating the maximum discrimination of an item at a particular value of the underlying trait; 3) the estimation of error of measurement at each ability level; and 4) and estimates that are not sample dependent (Crocker & Algina, 1986; Hambleton & Swaminathan, 1985). The added features and functionality of IRT are cited as benefits especially for measurement related tasks such as test development, scoring, scaling, equating, and item banking.

IRT models are also useful for assessing educational achievement and are currently used for such applications because of the additional measurement precision and flexibility. However, educational measurement is not the only domain in which IRT models have been applied. The measurement of attitudes, personality, opinions, psychopathology, and other non-cognitive traits have made use of sophisticated IRT models and have taken advantage of the benefits of IRT measurement. There is a variety of parameterizations or specifications of IRT models depending upon the theories about the construct being measured, the observed data, the assumptions that underlie those data, types of items, scoring of items, sample size, and number of items.

Thurstone (1927, 1928) and Likert (1932) can be credited for their contributions to the measurement of attitudes, opinions, and other non-cognitive constructs. Both Thurstone and Likert developed methods for scale development for the purpose of measuring non-cognitive traits, though their methods differ in important ways, namely in the assumptions about how people respond to items and criteria for item selection during the scale construction process. Likert’s (1932) method is much more widely used in practice, perhaps because it is less laborious. Although CTT approaches were initially applied to scale development for cognitive traits, they are also commonly implemented for non-cognitive measurement. These approaches are more congruent with the Likert methodology for measurement. Despite the fact that cumulative IRT methods were initially developed for the measurement of cognitive traits, they have been used for the measurement of personality, opinion, satisfaction and other non-cognitive latent traits. A specific class of models, called cumulative IRT models are also appropriately applied within the context of Likert approach to scale development and scoring. There are, however, fundamental differences between the assumptions of Likert’s and Thurstone’s (1927, 1928) approach to the measurement of non-cognitive traits, which would necessarily influence the consideration of model selection.

Although the introduction and application of IRT models have helped to advance the field of cognitive and non-cognitive measurement, a disparity seems to exist between the assumptions underlying non-cognitive data and some of the classes of IRT models that are typically applied to such data. Specifically, the IRT models that are commonly applied to non-cognitive data are probabilistic models defined by monotonic increasing functions (Andrich, 1996). The premise behind cumulative IRT models is that the probability of endorsement (or correct answer) of an item measuring a particular construct increases as the level of the construct an individual possesses increases (Andrich, 1996). Models with such characteristics are known as cumulative models.

In contrast, data resulting from the measurement of non-cognitive concepts such as personality, attitude, opinion, or satisfaction, have been called “unfolding” data. In unfolding-type data, direction is not always implied in observed responses to the agree/disagree items typically found of instruments used to measure such concepts, these data are considered “folded.” In order to appropriately connect these data to the concepts they are intended to measure, the data therefore must be ”unfolded” where Coombs (1950, 1964) used the term unfolding to describe the process of ascertaining the direction of the scale from observed responses.

Folding occurs especially for neutral items (Roberts, Laughlin, & Wedell, 1999; Stark, Chernyshenko, Drasgow, & Williams, 2006). For example, an individual asked to rate level of agreement with the following item eliciting an opinion about abortion would exhibit folding because a response of, say, ‘Disagree’, would not necessarily indicate that individual’s position: “Sometimes I am in favor of a woman’s right to abortion, but at other times I am not (see Roberts et al., 1999, p. 217). An item with a more extreme sentiment, in either direction, would probably not exhibit folding, as in the case of the following item: “Society has no right to limit a woman’s access to abortion” (Roberts et al., 1999, p. 217). An observed response of ‘Disagree’ with this item would provide an indication of an individual’s attitude towards abortion. Cumulative IRT models, by design, cannot accurately describe folded data. Unfolding IRT models are however specifically designed to measure data for which direction is not necessarily implicit.

The application of IRT unfolding models is most appropriate when data are folded or of the unfolding-type. IRT unfolding models for non-cognitive measurement are not new to the field of measurement where Thurstone’s (1927, 1928) work alluded to such approaches in his efforts towards measuring attitudes. Coombs (1950, 1964) derived the first unfolding model within the context of non-cognitive measurement, formally coined the term “unfolding” and described the process used by respondents when rating level of agreement with items that measure constructs like opinion, preference, or attitude. The presumed response process in operation is fundamentally different between an item measuring attitudes (i.e., non-cognitive) and an item measuring achievement (i.e., cognitive).

The distance between the location of an item and an individual on the latent trait continuum is the focal point of unfolding models. The premise of these models is that the probability of endorsing an item increases as the distance between the item and the individual on the latent trait continuum decreases. According to Roberts et al. “unfolding models operate on the basis of the absolute distance between an individual and an item on the continuum. . .” (1999, p. 213). The use of unfolding models, however, has not been as widespread as that of cumulative models, presumably because they are more complex (Stark et al., 2006). Additionally, unfolding IRT models have not endured a lengthy history of rigorous testing and evaluation that cumulative IRT models have. As a result, applied research using unfolding models is in its relatively early stage, compared to applied research using cumulative IRT models.

Consideration and use of unfolding models warrants attention to address a methodological issue that has been found to exist in applied research. The problem, which could lead to unintended and undesirable consequences is the lack of congruence between the assumptions that underlie the cumulative IRT models and those that underlie unfolding data. This can lead to inaccurate results. When decisions are made based on test scores any inaccuracies in scores can be problematic. For example, vocational hiring decisions are sometimes made based on personality instrument scores (Stark et al., 2006); surveys of satisfaction are commonly administered to recipients of services by an agency for the purpose of program evaluation, where decisions are made about the quality of and about efforts for improving or revising such programs. Further investigation into non-cognitive test development, scoring, and analyses is worthwhile because of the added accuracy of measurement, and validity of test score use they can afford in some non-cognitive measurement situations.

This study contributes to the body of empirical research concerning the application of unfolding models to non-cognitive data and provides insight into methods for non-cognitive scale construction and IRT model selection. Methods for examining the differences between cumulative and unfolding, unidimensional, parametric models as applied to Likert-type data are employed. The data result from a survey eliciting teachers’ perceptions of their working conditions. Examination of person and item locations on two separate latent traits (in the current study, teachers’ perceptions of leadership in their school and perceptions of teacher empowerment)  are made across the different scaling methods. Measures of statistical and graphical model-data fit are presented for each of the scoring methods and for each of the two constructs. Relative comparisons of model-data fit across the scoring models are made using the information theory based statistics Akaike Information Criterion (AIC; Akaike, 1974) and Bayes Information Criterion (BIC; Schwarz, 1978). These can be used as measures of fit across non-nested models, while taking into consideration the varying number of independently adjusted parameters, or the number of parameters to be estimated.

Finally, other properties of the statistical probabilities of responses to items across methods are examined to inform decisions about model-data fit. Data come from the 2006 administration of the North Carolina Teacher Working Conditions (NCTWC) survey. An example Likert item from the NCTWCS reads: “Teachers are held to high professional standards for delivering instruction.” Respondents are instructed to rate their level of agreement with items and response options include: strongly disagree, disagree, neither agree nor disagree, agree, and strongly agree. Data from the NCTWCS are used by government agencies for the purpose of informing educational policy, local school districts, principals and teachers for the purpose of improving teacher working conditions for attainment of the ultimate goal of improving student learning.

## Purpose of Study

The primary purpose of this investigation is to contribute to efforts for the advancement and improvement of methodologies for the measurement of non-cognitive constructs. In some measurement situations, unfolding IRT models can offer greater flexibility and precision over the traditionally employed measurement models such as CTT or cumulative IRT models. Analytically, person and item estimates from the application of cumulative and unfolding, unidimensional, parametric IRT models to real, attitudinal survey data are examined and compared in order to determine the feasibility and flexibility of the application of unfolding IRT models to such data. Two cumulative IRT models for polytomous data are employed: Masters' (1982) partial credit model (PCM) and Muraki's (1992) generalized partial credit (GPCM) model. The unfolding model used is the parameterization of the generalized graded unfolding model (GGUM) developed by Roberts et al. (2000). Additionally, a fourth scaling method, a structural equation model (SEM)—a commonly implemented procedure in applied, non-cognitive measurement situations—is also included.

A secondary purpose of this investigation includes the examination of fit of the generalized graded unfolding model to the data compared to the other two IRT models using graphical and statistical techniques. Chernyshenko et al. (2001) argued that more attention should be given to the fit of a particular model to data for the purpose of accurate interpretation of results. Their argument for the importance of assessing the accuracy of model-data fit is based on the fact that important decisions are often based on item/test analyses and results. Not all researchers share this sentiment, however. Common measures of model-data fit are often in the form of a chi-square statistic, which is sensitive to sample size and number of parameters to be estimated in the model (Hambleton & Swaminathan, 1985; Roberts et al., 2000). Alternatively, graphical representations of model and item fit have been described by Hambleton and Swaminathan (1985) and Roberts et al. (2000). For example, comparisons can be made between expected and observed responses with respect to $\hat{\theta}_j - \hat{\delta}_i$, or the difference (i.e., distance) between person and item estimates, which are measured on the same underlying latent trait scale. Roberts et al. (2000) used a similar methodology in that for every item-person pair, these differences were distributed into equally-sized, homogeneous groups. Average observed and expected responses were calculated for each group and then plotted against the average $\hat{\theta}_j - \hat{\delta}_i$ for each group. This approach is not a probabilistic one, but a graphical approach for examining item misfit (Roberts et al., 2000). Hambleton and Swaminathan (1985) described several methods to examine model-data fit, one which includes evaluation of residuals against $\hat{\theta}_j - \hat{\delta}_i$, at the item, person, or test level.

Due to the limitations of the chi-square statistic including the inhibition of relative comparisons across non-nested models and the dependency of the hypothesis test on sample size (i.e., any IRT model will be rejected with a large sample), a potentially more useful and informative statistic will be calculated as a measure of fit that is comparable across the three IRT models. Two examples of information theory-based statistics are the AIC and BIC. These indices are sensitive to “over fitting,” thereby favoring simpler models (Kline, 2005). Both are appropriate for comparing fit across non-nested models with varying number of parameters when maximum likelihood methods of estimating model parameters are used (Kang & Cohen, 2007). The BIC differs from the AIC in that the former directly considers the sample size by “penalizing overparameterization with the use of a logarithmic function of sample size” (Kang & Cohen, 2007, p. 333) and generally penalizes models more than the AIC when the sample is large (Bozdogan, 1987). The information-based statistics or criteria do not have known distributions, thus significance tests are not possible, as they are when using a statistic that is chi-square distributed (Kang & Cohen, 2007). As a result, “comparisons are made based on relative magnitude” (Kang & Cohen, 2007, p. 332), where the model with the smallest AIC or BIC is selected over competing models.  In addition to information theory-based criteria, graphical depictions of model/item-data fit for the three IRT models are constructed and compared. Specifically, the average observed item scores are plotted with the respective model-based predicted average item scores.

Although not a direct test of model fit, a component of the model-fit examination should include tests of assumptions of the models (Chernyshenko et al., 2001; Chernyshenko et al., 2007; Hambleton & Swaminathan, 1985). There are many methods of assessing test dimensionality within the framework of cumulative IRT models. Methods for testing the assumption of unidimensionality for unfolding data do not exist to the extent that their cumulative counterparts do (Habing, Finch, & Roberts, 2005; Stark et al., 2006). Davison (1977) described and illustrated the correlational and factor structure of unidimensional unfolding data, where Maraun and Rossi (2001) and van Schuur and Kiers (1994) further explained the structure of such data and the statistical consequences of applying linear factor analytic methods to unfolding data. Habing et al. (2005) appear to be the first researchers to derive a method for statistically testing the hypothesis of unidimensionality for data of the unfolding type. They suggested a modified version of Yen's (1984, 1993) $Q_3$ statistic specifically as a method to assess the assumption of unidimensionality in unfolding data that conform to the GGUM (i.e., observed graded-response data that describe level of agreement). The data and initial parameter estimates used in Habing et al. (2005) came from Roberts et al. (2000). The primary reason that the Habing et al. (2005) modified $Q_3$ statistic cannot be exploited in the current investigation is due to the fact that the GGUM  was previously determined by Roberts et al., (2002) to fit the observed data; a conclusion that cannot be made in this investigation. Examination of the dimensionality structure of the data in this investigation is implemented through the application of factor analytic methods.

Finally, to simultaneously examine fit and test the model assumption of local independence, chi-square statistics for item pairs and triplets (Chernyshenko et al., 2001; Stark et al., 2006) can be used to asses fit between each model and the data. Calculation of chi-square distributed fit statistics is a common approach to examine goodness-of-fit, although some caution must be exercised because such statistics are sensitive to sample size. Chernyskenko et al. (2001) and Stark et al. (2006) reported that typical chi-square statistics for each item may not be completely accurate because they are not necessarily affected by violations of other assumptions of both cumulative and unfolding data (i.e., local item independence and unidimensionality). Also, when using a one-parameter IRT model, the chi-square statistic may be unreliable if all items are indeed not equally discriminating (Chernyshenko et al., 2001). Thus, Chernyshenko et al. (2001) computed adjusted (to degrees of freedom) chi-square statistics for item pairs and triplets, which is an approach that is executed in this study for the purpose of gaining evidence for the appropriateness of the application of the three IRT models (i.e., PCM, GPCM, and GGUM).

## Research Questions

This empirical study focused on the application of both traditional and relatively new IRT measurement models to real attitudinal survey data. The primary purpose of this investigation is to examine differences in item parameters estimated by three IRT models (i.e., PCM, GPCM, and GGUM) and differences in person parameters estimated across those models and one SEM (i.e., confirmatory factor analysis) measurement model to provide insight into how both scale construction and construct measurement might be changed or improved. Graphical and statistical fit of models function as the secondary purpose of this study. The following four research questions are posed to address these purposes:

(1) Do the three IRT methods of scaling and scoring differ in terms of the ordering of item parameters from an attitudinal measure?

(2) How do the IRT and SEM methods of scaling and scoring compare in terms of the ordering of person parameters/estimates on the underlying latent trait of the attitudinal measure?

(3) Is the assumption tenable that the responses to the items on the attitudinal measure follow from an ideal point response process resulting in single-peaked, non-monotonic item characteristic curves? In other words, do item responses follow an unfolding pattern?

(4) How does the generalized graded unfolding model compare in terms of model-data fit with the partial credit and generalized partial credit models?

## Summary

As decisions continue to be made based on measures of personality, attitude, opinion, satisfaction, or preference, the need exists for reliable methods for scale construction and accurate estimates of responses to such measures. Methods for building a survey or test and the psychology of survey response must be considered simultaneously and exist in concordance with one another. Choice of model selection or method of scoring/scaling is also not made separately and independently from all other steps in the measurement process. Assumptions that define methods of scale construction, the item response process, and measurement model must ideally be aligned. Unintended consequences of dissonance between assumptions can yield inaccurate results, leading to ill-informed decisions.

Non-cognitive measurement is and has been conducted within a variety of disciplines where often times high stakes decisions are made. The assessment of personality is conducted for diagnostic and assessment purposes as well as for rehabilitation plans. Additionally, employers use personality testing to make hiring and other vocational decisions. Marketing research frequently use surveys to measure preference for or satisfaction with particular products; surveys are also used as a tool for evaluating social service or educational programs by eliciting attitudes, opinions, and satisfaction with services. The current investigation used data from the North Carolina Teacher Working Conditions Survey. This survey data is used by policy makers to change or institute policies that determine how administrators are educated and trained. Additionally, the way in which schools are funded rest partially on the survey results.

Within the context of non-cognitive measurement where respondents are asked to rate their level of agreement with an item, it is argued that the probability of item endorsement is high to the extent that the content and/or sentiment of that item closely matches the sentiment of the individual. Application of an unfolding IRT model would be a practical approach for scaling when this response process is responsible for producing the observed data. Unfolding IRT models are flexible, by design, in that they can accommodate the scaling of a continuum of item sentiment ranging from extremely negatively, to neutrally, to extremely positively worded items. Additionally, unfolding IRT models could prove efficient when items that comprise a survey that span the entire spectrum of the latent trait, from negative to positive. Investigation into an arguably feasible, alternative approach for measuring non-cognitive constructs seems warranted given the flexibility of unfolding models in scaling data, the capability of informing the scale construction process, and in some measurement situations the superiority over cumulative IRT models in measurement precision.

# CHAPTER 2

# LITERATURE REVIEW

A variety of latent trait models have been developed within the context of item response theory (IRT). These measurement models are not new, with some of the earliest work and contributions to this theory of measurement dating from the 1930s (Hambleton & Swaminathan, 1985). These models are often applied to cognitive data like achievement test data or attitudinal, behavioral, personality or other non-cognitive data. Briefly, IRT models are mathematical models that permit prediction of examinee test performance from an individual’s standing on an attribute or trait and the characteristics of the items that make up a test (Hambleton & Swaminathan, 1985). The relationship between observed performance on an individual item or total test and the latent trait of the examinee is specified by a particular IRT model and thus by the item characteristic curve (ICC) (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan & Rogers, 1991). Although there are numerous IRT models available for estimating person (i.e., latent trait, usually ability) and item parameters, all unidimensional IRT models rely on the assumptions of unidimensionality, local independence, and a monotonic (increasing) relationship between the probability of answering an item correctly (or endorsing an item) and the latent trait being measured (Hambleton et al., 1991; Hambleton & Swaminathan, 1985). This monotonic increasing relationship describes the shape of the ICC. The assumption of unidimensionality means that only a single latent trait (i.e., ability) sufficiently predicts examinees’ test performance, or that only one construct is being measured and only one construct explains examinee performance (Hambleton & Swaminathan, 1985; Hambleton et al., 1991). In practice, this assumption does not hold in a strict sense;  the unidimensionality assumption is considered to be satisfied when a single primary trait accounts for test performance and when the relationship between the underlying trait and test performance is the same for all subpopulations of test takers (Hambleton & Swaminathan, 1985). The assumption that a person's response to or performance on one item is not dependent on a response to a different item, when ability is held constant, describes local independence. In other words, responses to individual items are statistically independent from one another when conditioned on ability (Hambleton & Swaminathan, 1985; Hambleton et al., 1991). The nonlinear relationship between the latent trait and performance on an item is given by the mathematical function called the item response function (IRF). The difference between the IRF and the ICC is slight: the IRF is the mathematical equation or function; the ICC is just the graphical plot of the IRF. The assumption that the ICC is always increasing describes the monotonicity assumption. In cumulative IRT models, then, it is always assumed that as the level of the latent trait increases, so does the probability of endorsing (for non-cognitive items) or answering correctly (for cognitive items). IRT models differ in terms of the number of item parameters to be estimated, the scoring of different item types (i.e., a dichotomously-scored multiple choice item, a polytomously-scored short answer item where partial credit is granted, a polytomously scored Likert-type item with multiple ordered response options) and the number of latent traits that explain test performance.

### *Cumulative IRT Models*

Some of the more commonly applied parametric, unidimensional IRT models for dichotomously scored data are known as cumulative IRT models. They include: the Rasch, or 1-parameter logistic (1PL) model originally developed by Rasch (as cited in Smith & Smith, 2004; Yen & Fitzpatrick, 2004); the two-parameter logistic (2PL) model developed by Birnbaum in the 1950s (as cited in Hambleton & Swaminathan, 1985); and the three-parametric logistic (3PL) model (as cited in Hambleton & Swaminathan, 1985; Hambleton et al., 1991). Common unidimensional IRT models for polytomously-scored items include the following five approaches:

1) the partial credit model (PCM; Masters, 1982) which models the probability of successfully completing or responding to the k$^{th}$ item response category;

2) the generalized partial credit model (GPCM; Muraki, 1992) which is "formulated based on the assumption that the probability of selecting the k$^{th}$ category over the k minus first (k-1) category is governed by the dichotomous response model" (p. 160), and primarily differs from the PCM in that the discrimination item parameter is estimated in the GPCM;

3) the rating scale model (Andrich, 1978) which models the probability of selecting a particular ordered category where the same set of response options are associated with each item, that all items are assumed to have the same underlying thresholds for those response options, and that all items are equally discriminating (a feature of all Rasch or 1PL models) (Masters, 1982; Smith & Smith, 2004). Such a model would be appropriate for Likert-type items where all items have the same response options,

where it is assumed that all items discriminate equally, and that the item response options are used equally across all respondents for all items;

4) the nominal response model (Bock ,1972) which yields the probability that a person of a particular level of the latent trait will endorse or respond to a particular response category. This model has been described as a very general model (Yen & Fitzpatrick, 2004) where "item scores are in $m_i$ unordered categories and a higher item score does not necessarily reflect better performance" (p. 17). The nominal response model differs only slightly from the partial credit model and actually "becomes the PCM when the slope [or discrimination item] parameters are constrained to increase in steps of unity" (Thissen & Steinberg, 1986, p. 571); and

5) the graded response model (Samejima, 1969) for polytomously scored items with ordered response categories, which estimates the function that relates the latent trait to each score associated with each response option for each item. Specifically, "the probability of a response in category k or above" (Thissen & Steinberg, 1986, p. 569) is estimated by the graded response model. That is, similar to the partial credit model, the graded response model estimates separate thresholds for each response option. The graded response model differs slightly from the rating scale model in that the latter assumes the thresholds for all response categories are equidistant and that they are the same for all items (Yen & Fitzpatrick, 2004).

All of the models just described are categorized as cumulative; the primary underlying assumption is that the probability that an individual will agree with, endorse, or answer correctly an item increases to the extent that that individual's standing on the latent trait dominates or is more positive or greater than the content of the item (Roberts & Laughlin, 1996).

### *Additional Models Appropriate for Attitudinal or Preference Data*

Although this study will only address unidimensional IRT models, it is important to note that there are also several multidimensional IRT models used when more than one latent trait is expected or required to endorse or successfully complete an item. Research regarding the development of the theory and application of such models can be found in Ackerman (1992, 1994, 1996), Luecht and Miller (1992), Mislevy and Verhelst (1990), and Reckase (1985). Additionally, there exist unidimensional nonparametric IRT models for dichotomous and polytomous data, where fewer assumptions are made about the relationship between the probability of item endorsement (or getting an item correct) and the underlying latent trait, and the assumptions that are made are less strict. For example, one assumption in Mokken's (1997) nonparametric model for dichotomous data is that of monotonic homogeneity, or that the probability of item endorsement is monotonic non-decreasing. This is less strict than the monotonicity assumption of parametric models. The section on nonparametric models in van der Linden and Hambleton (1997) provides a brief overview of three common nonparametric IRT models. Finally, social science researchers have used multidimensional scaling techniques to interpret ordinal or "pick any" data. Pick any data refers to data that result when an individual is asked to select any number of choices from a list based on some criterion. For example, people could be asked to select any number of qualities they value in a supervisor, from a list of say 15 descriptors. A company may use pick any data to assist in making marketing decisions with individuals being asked to select from a list of products that they would actually purchase.

Selecting the most appropriate measurement model depends on the purpose of the analysis, the sample (i.e., size, characteristics), the construct(s) that the total test and component items are purported to measure, the scoring of the items (i.e., dichotomous, polytomous), the scoring of the total test, other attributes of the items such as speededness and probability or opportunity for guessing, and the response process assumed to underlie observed data.

*Responses Processes for Cumulative Models*

The parametric, unidimensional, cumulative IRT models previously described share an important, underlying mechanism: the “cumulative (monotonic increasing probability) response function” (Andrich, 1996, p. 349) and the very closely related dominance response process assumed to underlie the observed data (Chernyshenko, Stark, Drasgow, & Roberts, 2007; Roberts & Laughlin, 1996; and Stark et al., 2006). According to Roberts et al.:

> In a dominance response process, an individual endorses an item to the extent that the individual is located above the item on the underlying continuum. Responses from a dominance process generally are analyzed with some form of cumulative model in which the probability of endorsement increases as the signed distance between the individual and the item on the attitude continuum increases (1999, p. 215).

Data assumed to be produced by a dominance response process, where cumulative models to examine such data are most appropriately applied, are usually associated with the Likert approach to scaling and attitude measurement (Andrich, 1996; Chernyshenko, et al., 2007; Luo, Andrich, & Styles, 1998; Roberts & Laughlin, 1996; Roberts et al., 1999; Stark et al., 2006).

*Unfolding IRT Models*

A different class of parametric, unidimensional IRT models for dichotomous and/or polytomous data are known as unfolding models and are sometimes referred to as models for nonmonotonic items (van der Linden & Hambleton, 1997). Some of the earliest contributions to the theory of such measurement can be traced back to the work of Thurstone (1927, 1928) and his methods of psychological, attitudinal scaling, which implied a nonmonotonic response process where individuals are least likely to respond positively or endorse an item as the distance between a person and an item on a latent trait continuum increases. Thus, according to Thurstone (1927, 1928) people are more likely to endorse an item when the item most closely matches the attitude of the person (i.e., as the distance between a person and the item decreases). Coombs (1952, 1964) built upon the underlying theories of Thurstone (1927, 1928)--formally developed and termed unfolding models--and further developed the theory of data and measurement. Other researchers extended Coombs' (1952, 1964) deterministic models to define stochastic or probabilistic unfolding models (Andrich, 1988; Andrich, 1996; Andrich & Luo, 1993; Davison, 1977; Hoijtink, 1990, 1991; Roberts & Laughlin, 1996; Roberts, Donoghue, & Laughlin, 2000).

Thurstone's (1927) law of comparative judgment and proposed methods of measuring attitudes (1928) are credited by many scholars (e.g., Andrich, 1988; Andrich & Styles, 1998; Coombs, 1950; Stark et al., 2006) as some of the earliest contributions to the underlying theory of attitudinal or non-cognitive measurement. Thurstone's (1928) theory of attitudinal measurement states that a person's attitude is indicated by the selection of "a particular opinion which most nearly represents his own attitude" (p. 539). Essentially, "it was assumed that individuals would agree only with those statements that reflected their own attitude and would disagree with those statements that reflected either a more intense or less intense attitude than their own attitude" (Andrich & Styles, 1998, p. 455).

Coombs (1950, 1964) and Coombs and Avrunin (1979) built upon the idea of endorsing an attitude item only when the item closely matches that of the individual. Coombs's (1964) theory of data describes the relationship of data points, with a point being both an individual's agreement with or preference for or location in some space or continuum about a construct, as well as the location of the item in that same space. The location of the person in this space or on the construct continuum is what Coombs (1964) termed the person's *ideal point*. Thus, the ideal point process is an individual's response process and is in operation when an item that most closely matches the individual's attitude is endorsed (Coombs, 1964; Roberts et al., 1999).

One distinguishing characteristic between Coombs's (1964) and Thurstone's (1927, 1928) methods is that in the former, the location of individuals and items can be found concurrently (Andrich & Styles, 1998; Johnson & Junker, 2003; Noel, 1999). The idea of unfolding is closely related to the defining characteristics of the nonmonotonic, single-peaked response functions for non-cognitive data, and essentially refers to how the data are treated: they must be unfolded. Because a response of say, Strongly Disagree, is not informative in terms of the location of the person (i.e., above or below) in relation to the location of the item, data must be unfolded so that direction can be made explicit (Andrich, 1996). Roberts et al. (1999) "refer to the nonmonotonic behavior [of the ICC] as folding" (p. 216). This is why unfolding models are also referred to as proximity models; they describe that the "probability of endorsement is a function of the proximity between an individual and an item on the underlying attitude continuum" (Roberts et al., 1999, p. 213).

One limitation of Coombs's (1950, 1964) unfolding or proximity model is that it is deterministic or non-probabilistic (Andrich, 1988; Andrich & Luo, 1993; Andrich & Styles, 1998; Hoijtink, 1991) such that the probability of an individual endorsing an item can only be equal to 0 or 1 (Johnson & Junker, 2003). Coombs's (1964) model is of the form:

$X_{ai}=1$ if $|\beta_a-\delta_i| \leq \tau$ ,

$X_{ai}=0$ if $|\beta_a-\delta_i| > \tau$

where:

$X_{ai}$= the response to item $_i$ by person $_a$ ;

$\beta_a$= the location of person $_a$ on the latent trait continuum;

$\delta_i$= the location of item $_i$ ; and

$\tau$ = the "threshold (of equal size for each item) governing the maximum distance between $\beta_a$ and $\delta_i$ for which a person still renders a positive response" (Hoijtink, 1991, p. 154).

A probabilistic formulation of the item response function (IRF) of Coombs's (1964) model is specified as follows:

$P(X_{ai} = x_{ai}| \beta_a, \delta_i) = f(| \beta_a - \delta_i |, x_{ai})$ (Hoijtink, 1991).

The benefit of a probabilistic model is that stochastic parameterizations of the mathematical item response function $P(\theta)$ allow for statistical inferences to be made about person and item parameters (Johnson & Junker, 2003). Some parametric, probabilistic unfolding models developed recently include: Andrich's (1998) squared simple logistic model (SSLM); a hyperbolic cosine latent trait model (HCM) for dichotomous data (Andrich & Luo, 1993); a hyperbolic cosine latent trait model for polytomous data (Andrich, 1996); a generalized, reparameterized form of the hyperbolic cosine model (Luo, 1998); the parallelogram analysis (PARELLA) model for dichotomous data (Hoijtink, 1990, 1991); the graded unfolding model (GUM) for dichotomous or polytomous data (Roberts & Laughlin, 1996); and the generalized graded unfolding model (GGUM) for dichotomous or polytomous data (Roberts et al., 2000). Additionally, two nonparametric, probabilistic unfolding models for either dichotomous or polytomous data have been developed: the MUDFOLD model by van Schuur (as cited in Roberts et al., 1999), and an ordinal scaling method (Cliff, Collins, Zatkin, Gallipeau, & McCormick, 1988). Detailed information on the form and specification of parametric unfolding models is provided in a subsequent section of this chapter entitled "Specifics of Unfolding Models."

In theorizing how people respond to attitudinal items, Thurstone (1927, 1928) suggested that people only tend to endorse items that most closely match their perspective. Coombs (1964) further claimed that people agree with items that match their attitude/opinion, and disagree with item that contains a very different perspective, in either direction. The two theories share an important point; that is, a single peaked, as opposed to a "cumulative sigmoid shape[d]" (Andrich, 1988, p. 33) response function explains the relationship between item responses and the latent trait. To illustrate, consider the following statement with response options that range from strongly disagree to strongly agree (the number of response options is not important for this example): "This state's juvenile justice system treats criminals fairly." In response to this statement, 'strongly disagree' can be given for two reasons: that punishment is too harsh or too lenient. For a relatively neutrally worded item

> if the item is located far below the person's position on the attitude continuum (i.e., the item's content is much more negative than the person's attitude), then the person

> strongly disagrees from above the item. In contrast, if the item is located far above the person's position (i.e., the item's content is much more positive than the person's attitude), then the person strongly disagrees from below the item (Roberts et al., 2000, p. 4).

*Differences Between Cumulative and Unfolding IRT Models*

Although dominance and unfolding are both considered IRT models, they do differ conceptually and structurally. The more traditional IRT models were originally developed for test items within the framework of educational measurement where items measured underlying cognitive process like achievement or ability (Chernyshenko et al., 2007; van der Linden & Hambleton, 1997). On these types of tests, there is presumably one correct answer to each item, so the application of a cumulative model, with its respective underlying assumption of a dominance response process and monotonic increasing item characteristic curve makes sense for tests designed to measure knowledge, skill, or ability (Chernyshenko et al., 2007; Roberts et al., 1999; Stark et al., 2006; van der Linden & Hambleton, 1997). On tests that measure non-cognitive traits like attitude, satisfaction, or personality, one correct answer does not exist, so assumptions of monotonicity that describe cumulative response functions of dominance models would not hold.

To illustrate the structural differences between cumulative and unfolding models, the distinction between an item characteristic curve for a cumulative IRT model and an unfolding model can be seen in Figure 2.1 and Figure 2.2. Figure 1 shows a hypothetical item characteristic curve within a cumulative IRT framework. Figure 2.2 depicts an item characteristic curve from the perspective of an unfolding model.

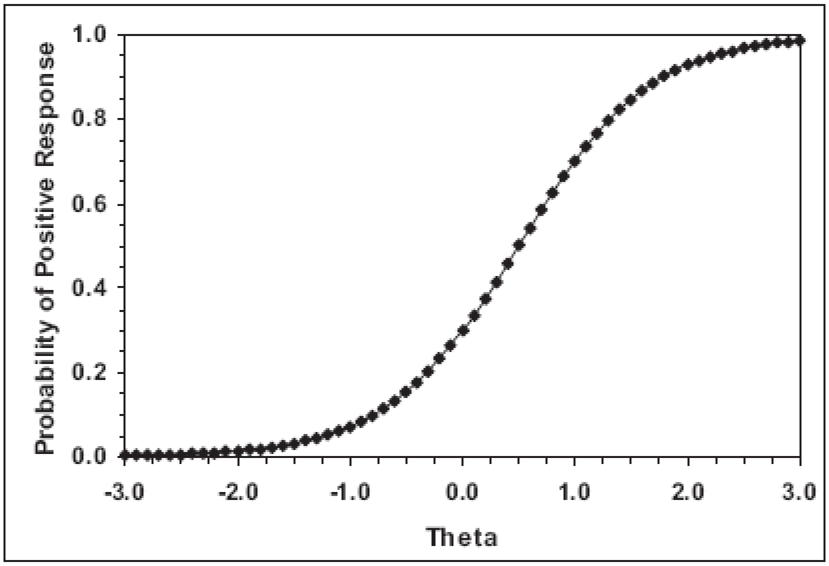


Figure 2.1. Hypothetical monotonic increasing Item Characteristic Curve (ICC) for a cumulative IRT model.

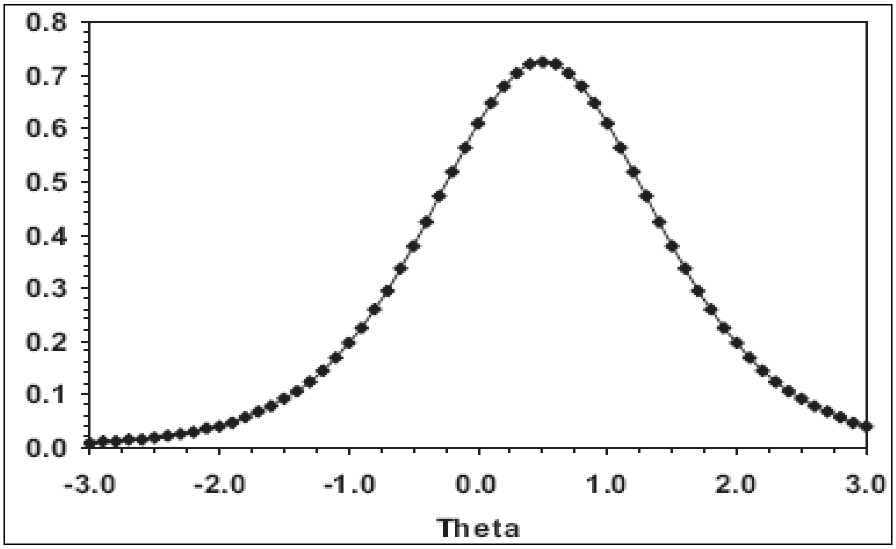


Figure 2.2. Hypothetical single-peaked Item Characteristic Curve (ICC) for an unfolding IRT model.

In cumulative models (see Figure 2.1), it is assumed that as the underlying trait increases, so does the probability of endorsing the item. In unfolding models (Figure 2.2), also known as proximity models, the difference in distance between the person and the item on the underlying trait describes the horizontal axis; thus as the distance between the two increase, the smaller the probability of endorsing the item.

Chernyshenko et al. (2007) argued that item analysis results from the application of ideal point process models contain information related to content, where as item parameters from cumulative models do not. For example, the *b*, or item difficulty parameter, in cumulative IRT models describes the difficulty of the item and influences the location of the ICC. Further, for the 1PL and 2PL models, the *b* parameter represents the point on the latent trait (usually denoted $\theta$) of a 50% probability of correctly answering or endorsing the item (Hambleton & Swaminathan, 1985). The higher the value of the *b* parameter, the more difficult the item. In unfolding models, however, item location, denoted $\delta_i$, represents the point on the latent trait where the probability of endorsing an item is greatest, and not just 50% (Chernyshenko et al., 2007). Table 2.1 provides a summary of the important differences between unidimensional, parametric, IRT cumulative and unfolding models.

The conceptual and structural differences between cumulative IRT and unfolding IRT models include the underlying response processes, the specification of the models, and the shape of the item characteristic curves. Another feature of unfolding models, distinguishable from cumulative models, is the distinction that must be made between the observed and unobserved response. As alluded to previously, a person may disagree with an item for one of two reasons, although which reason is not immediately known from the observed response. Parameterizations of unfolding models explicitly consider the lack of direction in observed responses where a respondent may disagree with an item because the sentiment is too extreme or not extreme enough. The following sections of this chapter provide descriptions of the more contemporary parametric, unfolding IRT models.

### *Unfolding Models Specified*

Andrich and Luo’s (1993; see also Andrich & Styles, 1998) hyperbolic cosine model for unfolding dichotomous data assumes that respondents can disagree for two reasons, from above or below, and agree for only one reason, thus yielding a model for three ordered, latent, response categories for two possible observed responses. Andrich (1996) developed a generalized hyperbolic cosine model for polytomous data, used with items having ordered response options such as: Strongly Disagree, Disagree, Agree, and Strongly Agree. This model assumes people can strongly disagree, disagree, and agree from above or below and only strongly agree for one reason (Andrich, 1996). This latter point results from the assumption that “the response of SA [strongly agree] implies that the person’s location is close to that of the statement” (Andrich, 1996, p. 353). In other words, a response of strongly agree implies a direction, whereas all other response options do not. Although the underlying theory is basically the same, the graded unfolding model (GUM; Roberts & Laughlin, 1996) and the generalized graded unfolding model (GGUM; Roberts et al., 2000) are parameterized slightly differently than the hyperbolic cosine models in that there are always two latent responses, or subjective response categories (SRCs; Roberts et al., 2000) for each observed response category (ORCs; Roberts et al., 2000).

Table 2.1

Comparison of Unidimensional, Parametric, IRT Cumulative and Unfolding Models

| Characteristic | Cumulative IRT Models | Unfolding IRT Models |
|---|---|---|
| Assumptions[a] | Monotonic increasing relationship between the observed responses and underlying trait | Non-monotonic relationship between the observed responses and underlying trait |
| Item Category Response Functions | Category response functions are observed and single-peaked | SRC: Single-peaked, symmetric around $\theta_j – \delta_i$<br>ORC: Single-peaked, or bimodal symmetric around $\theta_j – \delta_i$ |
| Item Characteristic Curve | S-shaped, always increasing / Monotonic | Single-peaked / Non-monotonic |
| Underlying Response Process | Dominance | Ideal Point |
| Type of Model | Dominance | Proximity |
| Type of data most appropriate | Cognitive (i.e., achievement) | Non-cognitive (i.e., attitudinal, personality) |
| Estimated Item Parameters[b] | $a_i$, $b_{iv}$, $b_i$, $d_v$ | $a_i$, $\tau_{ik}$, $\delta_i$ |
| Appropriate Scaling Method | Likert scaling method | Thurstone scaling method |
| Item Information Function[c] | A single-peaked function with a maximum at $\theta_j = \delta_i$ | A bi-modal function, symmetric about $\theta_j - \delta_i$ with a maximum at $\mid \theta_j = \delta_i \mid > 0$<br>Equal to 0 at $\theta_j = \delta_i$ |

Notes:

[a] Both unfolding and cumulative models assume unidimensionality and local independence.

[b] The GPCM item parameters are noted in this table for the cumulative model, and the GGUM item parameters are noted for the unfolding model.

[c] The test information functions (TIFs) are calculated the same way in both unfolding and cumulative models (i.e., TIF is the sum of each item information function).

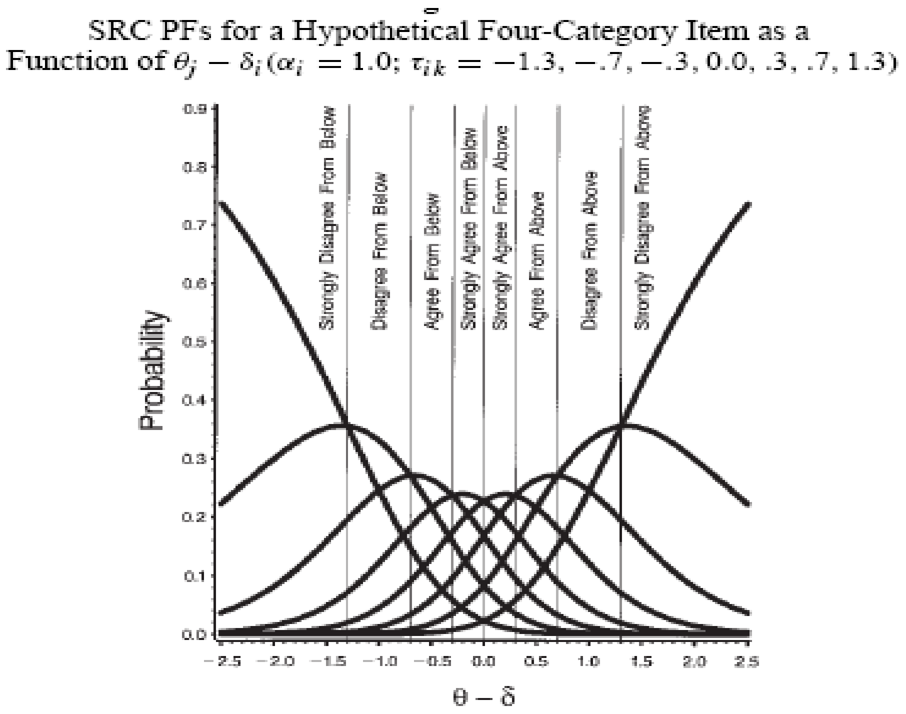


Figure 2.3. Plot of 8 Subject Response Category Probability Functions for a 4 Response Option Item (from Roberts et al., 2000, p. 5).

For example, Figure 2.3 depicts an item with four response options that has eight SRCs and seven threshold parameters (Roberts et al., 2000, p. 5). The item threshold parameter, $\tau_{ik}$, in the GUM, GGUM, HCM, and GHCM is defined as "the location of the k$^{th}$ SRC threshold on the attitude continuum relative to the location of the i$^{th}$ item" (Roberts et al., 2000, p. 5). Of note is the fact that the two subjective responses associated with a particular ORC are mutually exclusive. As such, for each item, summing the probabilities of the subject responses yields the probability that an individual will respond using a specific ORC (Roberts et al., 2000). Figure 2.4 shows the graph of the ORCs for the same item as in Figure 2.3 (Roberts et al., 2000 p. 7).

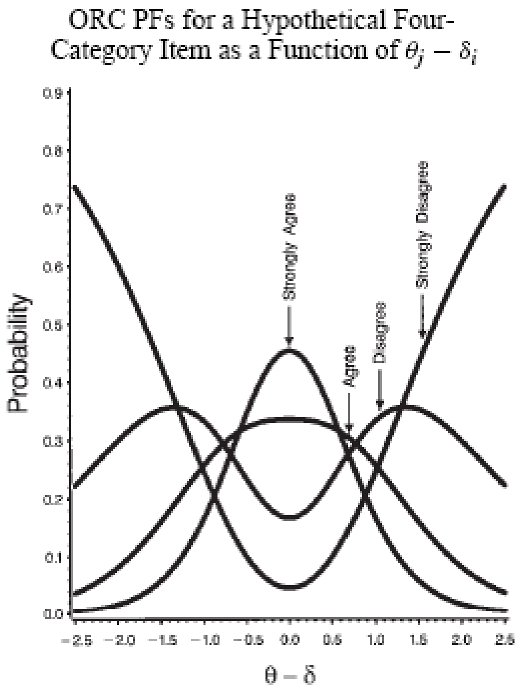


Figure 2.4. Plot of the 4 ORCs associated with the 8 SRCs in Figure 2.3 (from Roberts et al., 2000, p. 7).

Following Andrich and Luo (1993), Roberts and Laughlin (1996) modeled the subjective responses in their GUM with a cumulative IRT model, where they chose Andrich's (1978) rating scale model (Roberts & Laughlin, 1996). In the GGUM, Roberts et al. (2000) used Muraki's (1972) GPCM to model the subjective responses. Compared to the parameterization of the SRCs, in the GUM and GGUM parameterization of the probability of an observed response to a particular category of a particular item, the "$\tau_{ik}$ parameters lost their simple interpretation at the observable response level" (Roberts et al., 2000, p. 6).

The three equations for Muraki's (1993) GPCM, the GPCM as applied to the subject responses in the GGUM, and Roberts' et al. (2000) GGUM are shown in Equations (1), (2), and (3).

Muraki's (1993) GPCM is parameterized as follows:

$$P_{jk|k\text{-}1,k}(\theta) = \frac{P_{jk}(\theta)}{P_{j,k\text{-}1}(\theta) + P_{jk}(\theta)} = \frac{\exp[Da_j(\theta - b_{jk})]}{1 + \exp[Da_j(\theta - b_{jk})]} \quad , \qquad (1)$$

where k = 2,3,..$m_j$ and $m_j$ is the number of response categories.

The GPCM then is written as

$$P_{jk}(\theta) = \frac{\exp\left[\sum_{v=1}^{k} Z_{jv}(\theta)\right]}{\sum_{c=1}^{m_j} \exp\left[\sum_{v=1}^{c} Z_{jv}(\theta)\right]}$$

and

$Z_{jv}(\theta) = Da_j(\theta - b_{jv}) = Da_j(\theta - b_j + d_v)$ ,

where

D = a scaling constant (D=1.7);

$a_j$ = the item slope parameter;

$b_{jv}$ = the $v^{th}$ category parameter for item *j;*

$b_j$ = the item location parameter; and

$d_v$ = a category parameter.

Muraki's (1992, 1993) GPCM used by Roberts' et al. (2000) GGUM specifically to model the subjective response functions (SRFs) follows (Roberts et al., 2000, p. 4-5):

$$P(Y_i = y|\theta_j) = \frac{\exp\left\{ a_i \left[ y\,(\theta_j - \delta_i) - \sum_{k=0}^{y} \tau_{ik} \right] \right\}}{\sum_{w=0}^{M} \left\{ \exp\left\{ a_i \left[ w\,(\theta_j - \delta_i) - \sum_{k=0}^{w} \tau_{ik} \right] \right\} \right\}} , \qquad (2)$$

where

$Y_i$ = an observed response to item *i* ;

y = 0 (y = 0, 1, 2, . . . , M) corresponds to the strongest level of disagreement from below the item;

M = the number of subjective response categories (SRCs) minus 1;

$a_i$ = the discrimination of item *i*; and

$\tau_{ik}$ = the location of the $k^{\text{th}}$ SRC threshold on the attitude continuum relative to the location of the $i^{\text{th}}$ item.

The statistical parameterization of the GGUM (Roberts et al., 2000, p. 6) for observed responses is defined as follows:

$$P(Z_i = z)|\theta_j) = \frac{\exp\{ a_i [z (\theta_j - \delta_i) - \sum_{k=0}^{z} \tau_{ik}] \} + \exp\{ a_i [(M - z) (\theta_j - \delta_i) - \sum_{k=0}^{z} \tau_{ik}] \}}{\sum_{w=0}^{C} \{ \exp \{ a_i [w(\theta_j - \delta_i) - \sum_{k=0}^{w} \tau_{ik}] \} + \exp\{ a_i [(M - w) (\theta_j - \delta_i) - \sum_{k=0}^{w} \tau_{ik}] \}} \quad (3)$$

where

$Z_i$ = the observed response to statement *i* ;

*z* = is an index of agreement ranging from *z* = 0 to *C*, where 0 corresponds to the strongest level of disagreement and C corresponds to the strongest level of agreement; and

*C* = the number of ORCs minus 1 ($M = 2C + 1$).

With an explanation of unfolding models, assumptions that underlie them, and the parameterizations of models, specifically the parameterization of the generalized graded unfolding model (GGUM), the following section describes the need for and appropriateness of application of unfolding models to specific types of data (i.e., non-cognitive) and examines the literature as it relates to application of unfolding models to real and simulated data.

### *Applications of Unfolding Models for Theory and Model Development*

Proponents of the underlying theory of unfolding or ideal point process models have extended the theoretical work and applied such models to real and/or simulated data. For example, Andrich (1988) introduced the squared simple logistic model (SSLM) and applied it to both real and simulated dichotomous, direct-response data. The items used in Andrich (1988) elicited information about respondents' attitudes towards capital punishment. Similarly, Andrich and Luo (1993) compared analysis results like correlations and item and person parameters from the hyperbolic cosine model (HCM) and various parameterizations of it, and the SSLM using simulated data and the same real data as Andrich (1988). The focus of both studies was on the development of a new statistical model for the analysis of attitudinal data and on efforts to determine if estimates of the model parameters were possible. Andrich (1988) determined that the variances for both item and person parameters were always higher for the estimated parameters than for the initial or generating parameters, thus making estimates somewhat biased. Estimates for people and items from the SSLM were compared to existing Thurstone scaling estimates of the same instrument and were deemed to be equivalent (Andrich, 1988). A limitation of the SSLM, however, is such that the maximum probability of item endorsement is .5, even when the person and item have the same scale values.

Andrich and Luo (1993) were similarly motivated to derive a probabilistic model for unfolding data, though they made explicit the derivation of their general HCM, an unfolding model, from a cumulative model (i.e., the Rasch model for ordered response categories) and heavily emphasized one of the item parameters in their model, the unit parameter ($\theta_i$) described as a unit of measurement. Their focus was on unfolded models that needed to be folded in order to be compatible with folded data (Andrich & Luo, 1993). Folded data are those that result from Likert-type items where direction is not necessarily implied from the observed data, and can be easily recognized by a single-peaked response function. The two parameterizations of the HCM in Andrich and Luo (1993) were: the simple hyperbolic cosine model (SHCM) where the item parameter, called the unit parameter, denoted $\theta_i$, is held constant across all items; and the two-parameter hyperbolic cosine model (2PHCM) that allows the unit parameter to be estimated for each item. The difference between Andrich's (1988) SSLM and Andrich and Luo's (1993) SHCM is slight: in the SSLM "the square of ($\beta_n$ – $\delta_i$) is taken" whereas "the symmetric hyperbolic cosine of ($\beta_n$ – $\delta_i$) is taken" in the general and other parameterizations of the HMC (Andrich & Luo, 1993, p. 261). Andrich and Luo (1993) describe how the use of the SHCM over the SSLM is a substantive improvement where the SHCM "involves analyzing the details of the unfolding response mechanism to reveal and then model explicitly the two latent responses for a Disagree response, and then bringing them together" where the SSLM "is essentially based on a device (the square function) that produces a descriptive model of the required shape" (p. 261).

The characteristic that the two models share, however, is the maximum probability of item endorsement of .5. The simulation study of Andrich and Luo (1993) was conducted to determine item and person parameter recovery, which were deemed acceptable. Using real data (from Andrich, 1988) comparisons were made between SSLM (Andrich, 1988), SHCM and 2PHCM parameter estimates where the correlation between the SSLM and 2PHCM estimates was equal to .997 and to .999 between the SSLM and SHCM estimates (Andrich & Luo, 1993). Additionally the ordering of response patterns was compared for the SHCM and 2PHCM where only slight differences were found. Andrich and Luo (1993) did not make explicit comparisons across the three models for the person location, or $\beta_n$, parameters.

In related studies, Hoijtink (1990) introduced the probabilistic form of Coombs's model (1964), the parallelogram model, called PARELLA, for measuring latent traits (i.e., attitudes) and tested the model using simulations. Like Andrich (1988), Hoijtink (1990) intended to determine the feasibility of parameter estimates, though he used a marginal maximum likelihood (MML) approach (Bock & Aitkin, 1981) for item parameter estimates and the expected a posteriori (EAP; Bock & Aitkin, 1981) method for person estimates, whereas Andrich (1988) and Andrich and Luo (1993) used the joint maximum likelihood (JML; see Andrich & Luo, 1993 and Luo, 2000 for solution equations). Hoijtink's (1990) PARELLA model is similar to Andrich's (1988) SSLM though the models do differ in two ways. First, the former model involves an item parameter (in addition to item location) called the power parameter, denoted $\gamma$, and describes the importance (Hoijtink, 1990) of the distance between the person and item location on the latent trait. In the latter, an analogous item parameter, the unit parameter ($\theta_i$) is specified. Second, the maximum probability of the PARELLA model is 1.0 which is not true for the SSLM. Hoijtink (1990) specifically examined the stability and accuracy of estimates using simulations with variations of the following parameters: power parameter, shape of the person distribution, width of the person distribution, number of items, sample size, number of nodes (i.e., quadrature points), and the distribution of item parameters. A common theme within all results was the measure of the difference between the estimated and generating (initial) parameter estimates in Hoijtink

(1990) and in his follow up study (1991). A summary of the general trends reported by Hoijtink (1990) include: biased estimates of the person distribution (uniform and skewed) due to the restriction of range, although estimates were found to be unbiased for bi-modal and normal distributions. Recovery of the person distribution was not accurate for a sample of 100 people, but was for sample sizes of 300 and 900. Also, the increase from 10 to 16 nodes only improved the recovery of parameters for the person distribution for the condition of a bi-modal generating distribution. Hoijtink (1990) also employed a method for examining "fit" of the PARELLA model to the data using "the results of the E-step and the M-step to provide a way to evaluate the fit of the data to the expected stimulus characteristic curve" (p. 653) specifically using the difference between the observed and expected number of people at a particular quadrature point who selected item *i*. Although there is no criterion to evaluate such comparisons to determine statistical fit, this method appears to be an acceptable descriptive measure.

In his follow-up study using a simulated data set, Hoijtink (1991) again examined the appropriateness of the PARELLA model and examined model fit using the sum of the differences between the observed and expected outcomes related to the item response functions. The criteria used by Hoijtink (1991) to differentiate between good and poor fit are dependent upon on sample size and number of quadrature points, although for a sample of 300 people and 10 quadrature points "two or more differences between [empirical and expected values] greater than 4.0, or a sum of differences greater than 20 is indicative of an inequality between the empirical and expected IRFs" (p. 163). This criterion is only a rule of thumb, and Hoijtink (1991) recommended for a relaxation of the criterion for larger samples and a more stringent criterion for smaller samples. Hoijtink (1991) also applied the

PARELLA model to three real data sets, one of which was used in Andrich (1988) allowing some comparisons to be made with Andrich’s (1988) SSLM analyses. For example, the order of the item estimates was found to be the same across both PARELLA and SSLM models. Application of the PARELLA model was determined to have merit because of the replication of Andrich’s (1988) results, the exhibition of good fit according to the above criterion and examination of the weights associated with each quadrature point (the parallelogram structure of the data) (Hoijtink, 1991).

In an attempt to test the assumptions of the ideal point response process for the purpose of informing and improving personality assessment, Stark et al. (2006) applied two cumulative and two unfolding models to real personality data for the purpose of closely examining how people respond to personality items. Weekers and Meijer (2008) extended the work of Stark et al. (2006), although they used a similar methodology to compare analyses from cumulative and unfolding IRT models applied to surveys developed using dominance response process assumptions and ideal point response process assumptions. Chernyshenko et al. (2007) used a cumulative and unfolding model to score three surveys: one designed using ideal point response process assumptions using the GGUM, one using the traditional Likert approach (which, by design, assumes a dominance response process) to scale development using CTT, and another using dominance response process assumptions with a cumulative IRT model (2PL). Real data from a personality inventory were used and comparisons made between person and item parameters for the scale design (ideal point, dominance)/IRT statistical model (2PL, GGUM) pairs.

Stark et al. (2006) stated that it is not always apparent which response process (ideal point or dominance) is necessarily in operation and responsible for observed data from non-cognitive items. Thus, their purpose was to investigate if ideal point process models provide viable alternatives for scale development and scoring to the traditional and almost exclusively employed Likert-type or dominance IRT methodologies. Four IRT models—a parametric cumulative model (2PL), a nonparametric cumulative model  (Levine's nonparametric maximum likelihood formula scoring model, MFSM); a parametric ideal point model (GGUM),  and a nonparametric ideal point model (Levine's MFSM with ideal point constraints)–were applied to real, dichotomously-scored personality data comprised of 16 subscales, and directly compared in terms of chi-square fit statistics and graphical fit plots to determine the appropriateness of each model including the underlying assumptions (Stark et al., 2006). Because examination of the appropriateness of assumptions was a primary component to their investigation, Stark et al. (2006) calculated chi-square fit statistics for pairs of items and item triplets, as statistics for a single item are not necessarily sensitive to violations of the IRT assumptions of unidimensionality and local independence. The three chi-square statistics for each model were averaged across all items for each of the 16 subscales and directly compared across the four IRT models, where smaller chi-square statistics (< 3) was the criterion for good fit. The graphical representation of the ICCs from the four models was used as a second measure of fit, with specific focus on the extreme values of the underlying latent trait, as nonmonotonicity of ICCs is an indication of an ideal point response mechanism and it is in the extremes of the distribution where cumulative and unfolding models tend to diverge the most. Based on the two measures of fit, four of the 16 subscales were found to contain at least four nonmonotonic items. Of these four subscales, Stark et al. (2006) determined, based on the chi-square measure of fit, that the nonparametric ideal point model (i.e. Levine's MFSM with ideal point constraints) fit best for three, and the

MFSM with dominance constraints fit best for one subscale. The mere fact that any items exhibiting nonmonotonicity were found in addition to the fit measures were reason for Stark et al., (2006) to defend that ideal point models are flexible enough to model non-cognitive traits and should be considered as an alternative to cumulative or dominance models. This conclusion stems from the fact that the data used came from the *Sixteen Personality Factor Questionnaire* (*16PF)*, which was developed using a dominance response methodology (Stark et al., 2006).

Chernyshenko et al. (2007) pursued an investigation similar to that of Stark et al. (2006) by investigating the flexibility and functionality of a dominance model (2PL) and an ideal point model (GGUM). Each model was applied to three data sets produced by three scales; one built using traditional CTT methods, one using dominance IRT (2PL) methodology, and another using ideal point (GGUM) methodology.

Chernyshenko et al. compared IRT scores in two situations between the 2PL and the GGUM: scores compared when both models were applied to the data resulting from the scale built with the 2PL; and the comparison of scores when both models were applied to data resulting from the scale built with the GGUM. Chernyshenko et al. determined that the GGUM performed as well in fitting the 2PL data as the 2PL model, as evidenced by a .97 correlation between the IRT scores, however, results from the application of the 2PL model to GGUM data were not as favorable. Chernyshenko et al. showed that the item location parameters within the context of ideal point models provide information about item content, where those parameters from any dominance model are not related to item content. Additionally, correlations with four other measures were used for the purpose of providing evidence of criterion, discriminant, and convergent validity, where “the overall criterion-related validity of scores dropped in every case when 2PLM was used to score the Ideal Point order [construct measured] items but remained the same when GGUM was used to score the Dominance IRT order items" (p. 101).The general conclusion made by Chernyskenko et al. was that ideal point models should be considered and implemented as a mechanism to improve non-cognitive measurement. This decision was based on the fact that ideal point models can accurately model data resulting from dominance methodologies and they do provide added measurement precision, especially towards the middle of the latent trait by including items that contain neutral sentiments, making them much more flexible.

The study by Chernyshenko et al. (2007) built upon the investigation by Stark et al. (2006) in that the former constructed various scales according to the model assumptions, then applied the respective and alternative  models (i.e., applying the GGUM to 2PL data) to data, and made efforts to examine validity to empirically determine the flexibility and efficacy of ideal point models.  Weekers and Meijer (2008) attempted to replicate the findings of Chernyshenko et al. (2007), measuring a slightly different facet of the personality construct, and using different dominance and ideal point IRT models. Similar to Stark et al. (2006), Weekers and Meijer (2008) used four types of models (parametric, nonparametric, dominance, and ideal point) although the specific models differed. The parametric dominance model used by Weekers and Meijer (2008) was the 1PL, and the nonparametric dominance model used was Mokken's (1997) model of monotone homogeneity. The parametric unfolding model employed was the GGUM and the nonparametric multiple unidimensional unfolding model (MUDFOLD; van Schuur & Post, 1998, cited in Weekers & Meijer, 2008) was the nonparametric unfolding model used in the analyses. Like Chernyshenko et al. (2007), Weekers and Meijer applied a dominance and unfolding model to a data set resulting from a scale built from a dominance perspective and to another data set resulting from a scale build with ideal point assumptions. Weekers and Meijer found similar results in that the correlation between person estimates (i.e., IRT scores) was high for both the dominance scale (*r* = .988) and for the ideal point developed scale (*r* = .971), although they too found discrepancies between the unfolding and dominance model estimates for the unfolding data, especially at the upper extreme values of the underlying trait. Echoing the argument of Stark et al. (2006), Weekers and Meijer confirmed that inappropriate application of a model can have implications about decisions based on scores, as the ordering of people varied at the upper end of the trait between the dominance and unfolding models when applied to the unfolding data. As a result, inaccurate decisions would be made if the interest was focused on those people located at the upper 5% or 10% of the distribution. Finally, Weekers and Meijer drew similar conclusions to Chernyshenko et al. (2007) and Stark et al. (2006), in that they provided evidence in support of the use of unfolding models with non-cognitive data, not necessarily to replace cumulative (i.e., dominance) models, but to contribute to more precise measurement and improved scale development for certain non-cognitive constructs.

Although they calculated various chi-square fit statistics as one form of evidence in inform their decision making, a similar methodology was implemented across the Chernyshenko et al. (2007), Stark et al. (2006) and Weekers and Meijer (2008) studies; namely the use of a computer program MODFIT  to calculate adjusted chi-square fit statistics. These statistics, denoted $\chi^2/df$, are adjusted by dividing the value of the chi square statistic by their degrees of freedom. When building three types of scales using three methodologies, Chernyshenko et al. (2007) used MODFIT for both the 2PL and GGUM to calculate adjusted chi-square statistics for items, item pairs, and item triplets. These statistics were used to examine fit and to help determine which items from the larger item pool should be retained for the final 20-item scales. This comparison was facilitated by the equal sample size of 3,000. Weekers and Meijer (2008) used four different models, and four different software programs that each calculated a chi-square fit statistic, where MODFIT was used to calculate the adjusted chi-square fit statistics as a measure of GGUM model-data fit. Finally, comparisons of statistical fit, in a global sense, between the 2PL and the GGUM were made by Stark et al. (2006) as calculated with MODFIT, again using the same and equal sample size (i.e., 3,000) as Chernyshenko et al. (2007). The criterion used to evaluate comparisons is described in Chernyshenko et al. (2007) where it indicated that “previous studies have found that good model-data fit is associated with adjusted $\chi^2/df$ of 3 or less” (p. 93). (Chernyshenko et al. (2007, p. 93) present the derivation of the equation for a chi-square fit statistic for a single item is as follows:

$$\chi_i^2 = \sum_{u=0}^{1} \frac{[o_i(u) - E_i(u)]^2}{E_{i(u)}} \qquad (4)$$

where the expected frequency of selection of a particular response category for item *i* is calculated as:

$$E_i(u) = N \int P(U_i = u \mid \theta) * f(\theta) d\theta \qquad (5)$$

and, finally, the “the expected frequency for a chi-square statistic involving a pair of items, for items *i* and *i*’, in the (u,u’)th cell of a two-way contingency table, is computed as:

$$E_{i,i'}(u,u') = N\int P(U_i = u \mid \theta)P(U_{i'} = u' \mid \theta) * f(\theta)d\theta \,. \qquad (6)$$

To test the feasibility of their newly developed graded unfolding model (GUM), Roberts and Laughlin (1996) applied the model to simulated and real data pertaining to attitudes toward capital punishment. This model is an extension of Andrich and Luo’s (1993) hyperbolic cosine model (HCM) in that the GUM can accommodate either dichotomous or polytomous (i.e., graded) data, as opposed to dichotomous data only. Roberts and Laughlin (1996) developed the GUM with four guiding principles: 1) that an individual agrees with a statement to the extent that the statement reflects that individuals’ standing on the construct being measured; 2) an individual may agree or disagree with an item for two reasons (because the item does (not) express a strong enough sentiment, or that the item does (not) express too strong or too extreme of a sentiment) called agreeing or disagreeing from above or below; 3) the subjective responses (i.e., disagreeing/agreeing from above or below) are modeled using a cumulative IRT model; and 4) the subjective category thresholds (the point of intersection of the response functions for the subjective responses) “are symmetric about the point ($\theta_j - \delta_i$) = 0” (p. 235).

To examine the accuracy of the GUM estimates, Roberts and Laughlin (1996) used both real data and data generated from a simulation study with 30 conditions using the joint maximum likelihood method for parameter estimations. In the simulation study, they varied sample size with five values ranging from 100 to 1,000 and six values of test length ranging from five to 30 items where items had six response options. Item locations ($\delta_i$) were set equidistant, the thresholds (denoted $\tau_k$) or the point of intersection of the subjective response functions on the latent trait were held constant at .4, and the person parameters, theta ($\theta_j$), were randomly sampled from a normal distribution (Roberts & Laughlin, 1996). Four measures of accuracy were computed including the root mean square error (RMSE) for each of the three parameters, $\delta_i$, $\tau_k$, and $\theta_j$, a Pearson product moment correlation between the simulated (true) and estimated parameters, the ratio of estimated and true parameter variance, and the average difference between the estimated and true parameters (Roberts & Laughlin, 1996). Major findings included that accuracy of theta estimation is largely a function of the number of items, where accuracy increased with an increase in items. A similar effect was found with item location for the RMSE and variance ratio measure of accuracy and with threshold estimation for the RMSE, average difference and variance ratio measures (Roberts & Laughlin, 1996). Inaccuracies in parameter estimates were almost entirely a function of the difference in variance between the generating (true) and estimated parameters, which may have been due to the fact that the JML method of estimation was used (Roberts & Laughlin, 1996). Inaccuracies decreased as the number of items and people increased, though the decreased effect was more profound with the added items (Roberts & Laughlin, 1996). Overall, estimation was deemed possible and accurate by Roberts and Laughlin (1996) using the GUM on data that include at least 100 people with 15 to 20 items.

Roberts, Donoghue and Laughlin (1998) and Roberts et al. (2000) extended the work of Roberts and Laughlin (1996) to develop a probabilistic, unfolding IRT model for graded responses allowing both the item discrimination and item threshold parameters to vary across items, called the GGUM. Data were simulated in Roberts et al. (1998) to determine the accuracy of the MML item and EAP person estimates and to examine how well the GGUM recovered such parameters with six levels of test length ranging from five to 30 items (each with six response options) and six levels of sample size ranging from 200 to 2,000 people.

Briefly, Roberts et al. (1998) found that accurate item estimates resulted with at least a sample size of 750 and accurate person estimates resulted with at least 15 to 20 items. Roberts et al. (2000) described general, graphical methods to help identify items that fit poorly by applying the GGUM to real data about respondents' attitudes toward abortion where the final scale consisted of 20 items and the sample included 750 undergraduate students. Interesting findings from Roberts et al. (2000) include the differing effect of the two item parameters, discrimination ($a_i$) and threshold ($\tau_i$), on the expected value GGUM function and item information function. An increase in the discrimination parameter yields a more peaked expected value function that approaches its upper bound. The effect of increases in item thresholds (i.e., the distance between the subject responses relative to the location of the item) has a similar effect on the expected value function in that it approaches its upper bound, but this increase yields a flatter function (Roberts et al., 2000). According to Roberts et al. for item information functions:

> The information function becomes larger and more peaked as $a_i$ increases, but it becomes smaller and less peaked as $\psi$ [interthreshold distance] increases. Thus, maximum measurement precision is achieved at two symmetric points (or regions) on the latent continuum, and items with large discrimination indices and small interthreshold distances yield the most precision at these points (p. 17)

A final relationship found by Roberts et al. (2000) between the two item parameters, discrimination and threshold, was the quadratic relationship between the two estimates, where, as item locations tended toward the extremes, the absolute value of the threshold estimates increased. Although it is not clear if such a relationship can be generalized beyond their analysis, Roberts et al. (2000) interpreted this relationship to suggest "that moderate items distinguished among respondents more than extreme items" (p. 21).

Related to model fit, Roberts et al. (2000) calculated the difference between person estimates and item location estimates for all item (*n* = 20) and person (*n* = 750) pairs for a total of 15,000 differences. These differences were then grouped into 200 approximately homogeneous groups of size *n* = 75. Roberts et al. (2000) graphically plotted the average observed and expected responses, calculated Pearson product-moment correlations between these same sets of responses, and product-moment correlations between expected and observed responses for each item. Based on the overlap of the plotted scores and the high (*r* = .995) correlation between the observed and expected responses, Roberts et al. (2000) concluded that the global fit of the GGUM appeared to be reasonable.

### *Practical Application of Unfolding Models*

The use of unfolding models extends beyond simply testing the feasibility and capability of these models. Unfolding models have been applied to revisit research findings of the poor relationship between attitude and behavior as an alternative and potentially valuable way to better understand this relationship. Andrich and Styles  (1998) have argued that the poor relationship between attitude and behavior that exists in the extant literature "may be a methodological artifact that is related to the location of statements on a continuum as envisaged by Thurstone, and because Thurstone's methods are not used routinely in substantive research on attitude measurement…" (1998, p. 456).

The application of unfolding models as improved methods of understanding patterns of change and how people move or progress through developmental stages has received recent interest and can be found in the literature within a variety of contexts. The general argument for the superiority of unfolding models over cumulative models is such that single-peaked functions describe stage/developmental data better than monotonic increasing functions. Noel (1999) succinctly described this point:

> Obviously, such a developmental model relies on a different conceptualization of psychological change than do the better known cumulative models, such as Mokken, Rasch, or Guttman scaling. Whereas in cumulative models each stage is assumed to prepare the following in an integrative manner, so that earlier stages remain embedded in the later ones, in unfolding developmental models each stage is preparing the following while inhibiting the previous ones. Otherwise stated, the unfolding model of change assumes that some processes are relevant in a given stage but no longer relevant as one moves along the developmental continuum. (p. 175).

Noel (1999) applied Roberts and Laughlin’s (1996) GUM to test a proposed theory of cognitive and behavioral change as it related to a sample of cigarette smokers and their attitudes about smoking in an attempt to investigate the tenability of the proposed theory of change and the hypothesis that the change process is explained by a single-peaked pattern. The efficacy of such an inquiry could assist in treatment for smoking if reliable information exists about where people are in the various stages of change.

Another example of measuring progression through stages is provided in DeMars and Erwin (2003), where intellectual development of young adults was hypothesized to progress through stages. DeMars and Erwin’s (2003) investigation centered around a developmental instrument called the *Scale of Intellectual Development XII* (*SID-XII*), which was developed using the theoretical model postulating that adult development is stage-like. Their efforts were directed towards validating the underlying theoretical model (stage theory of adult development) upon which the *SID-XII* was built, in addition to producing a meaningful and informative single score on the latent trait, and to assess item fit (DeMars & Erwin, 2003). Motivation and implications of their study were couched within the context of higher education and improving assessment instruments and methodologies within higher education.

The overarching goal of a study by Touloumtzoglou (1999) was very similar to that of DeMars and Erwin (2003) in that the purpose was to examine the psychometric properties of a particular scale, elucidate observed scores from an ideal point response process perspective, and generally spawn efforts for improved measurement. Whereas DeMars and Erwin (2003) used the GGUM for polytomously scored items about intellectual development and allowed the discrimination of the items to vary, Touloumtzoglou (1999) employed the hyperbolic cosine model (HCM) on dichotomously scored items about attitude towards the visual arts.

Overall, each of the studies noted in the previous sections of this chapter provide support for the theoretical development and practical application of unfolding models. The structure of the theory that underlies unfolding models makes substantive sense and its application to some non-cognitive data seems appropriate. Empirical studies have provided evidence that application of unfolding models to non-cognitive data can indeed be superior to cumulative models in terms of model flexibility and improved measurement accuracy for extreme values of the trait, and more applied researchers are using unfolding models on real survey or attitudinal data. Additionally, the theory underpinning unfolding models has even prompted some researchers (i.e., Andrich, 1988; Andrich & Styles, 1998) to re-examine previous research findings from an ideal point perspective.

### *Improving Measurement with Unfolding Models*

Application of unfolding models to non-cognitive data seems reasonable because of the characteristics and theoretical components of the unfolding model described. The assumptions that underlie both the observed data and the unfolding model are more congruent than the application of cumulative models to unfolding data when items require a respondent to select a response category that most closely matches or reflects the person's attitude (Andrich, 1988, 1996; Chernyshenko et al., 2007; Noel, 1999; Roberts et al., 1999; Roberts et al., 2000; Stark et al., 2006). Unfortunately, there exists a discrepancy in current research methodologies between the assumptions of the underlying response process that produce the observed item scores, and how those items are developed and the scoring/scaling of the resulting data, where non-cognitive scales are still typically developed and scored using the traditional, Likert methodology (which implies a dominance response process). This is particularly problematic for attitudinal and personality data, as researchers have shown that the assumption of the ideal point response process (and the application of unfolding IRT models) better explains such data than the dominance response process (see, Andrich, 1996; Andrich & Styles, 1998; Chernyshenko et al., 2007; Roberts et al., 1999; Stark et al., 2006). Specifically, the problem lies in the interpretation of the results, which could be inaccurate if a model assuming a dominance response process is applied to unfolding data. The inaccuracies exist especially for individuals whose true location on the latent trait is extreme (Chernyshenko et al., 2007; Roberts et al., 1999) which can yield inaccurate decisions based on the results.

One of the fundamental differences between cumulative and unfolding models is the shape of the ICC, where monotonic increasing, ogive-shaped curves characterize the dominance response process, and single-peaked bell-shaped curves characterize the ideal point response process. Accordingly, there are always two person estimates that yield the same probability of item endorsement (Andrich, 1995a) within the context of unfolding models, and only one person location on the latent trait that is associated with a probability of a positive response when a cumulative model is applied. It is for this reason that closer examination and perhaps a change in the way researchers analyze non-cognitive data is warranted. If data are truly of the unfolding type, and a cumulative IRT model is applied, results will be inaccurate. The degree of inaccuracy depends on the relative location of the items and people.

In summary, Likert-type instruments are built, by design, to only include relatively positively and negatively worded items, have high internal consistency, and to have items with high item-total correlations. By design, relatively neutral items are omitted, thus omitting some level of measurement precision on the latent trait. Also, as a result of reverse scoring negative items, all ICCs should appear to be monotonic increasing. However, results from Meijer and Baneke (2004) in their analysis of data from the *Minnesota Multiphasic Personality Inventory - 2* (*MMPI*-2) and the analysis of data from the *Sixteen Personality Factor Scale* (*16PF*) by Chernyshenko et al. (2001) yielded items with nonmonotonic ICCs. Existence of some unexpected nonmonotonicty provides support for further consideration of unfolding models. Additional support for further investigation into the appropriateness of unfolding models for non-cognitive data was evidenced by Chernyshenko et al. (2007) who found that the conditional statistic, item information, for Likert-type items provided the most information towards the middle of the latent continuum. Thus, if a scale is built using the Likert methodology, items require a respondent to select a response category that most closely reflects the person's attitude, and users of the observed data are especially concerned with those sample members whose attitudes, opinions, perspectives are more extreme, then application of unfolding models would at least provide better measurement precision for respondents than their cumulative counterparts. In many practical situations, users of the data are most interested in those with extreme attitudes or perspectives. The ideal measurement situation would be such that survey development reflects the intended purpose of the survey and intended uses of the scores. Items contained on the survey should reflect the purpose, and assumed response processes that will govern the observed responses. Finally, analysis of the observed data would be very closely aligned with the assumptions made about the data.

Chernyshenko et al. (2007) emphasized the need for continued, improved personality measurement and clearly described the problems with commonly used methods. Their study focused on scale construction, which should be a first step because methods of, or procedures for, test development will dictate how data are used and analyzed. Chernyshenko et al. (2007) highlighted the problem and constraints with using classical test theory and factor analytic methods for scale construction and applying cumulative models for item analysis and scoring in the context of personality scale development. An explanation was detailed about the parallel between ideal point processes and unfolding models for test construction. Because of the greater flexibility that ideal point process or unfolding models offer, Chernyshenko et al. (2007) noted that:

> Constructing scales under ideal point assumptions would therefore allow the inclusion of items having a wider range of locations rather than just those tending toward

> extremes. This, in turn, would improve scale precision, reduce inventory development costs, and offer a relatively straightforward path toward computerized adaptive tests...” (p. 91).

Support for the claim that application of traditional, parametric, cumulative IRT models may not always be appropriate for non-cognitive survey development and data analysis comes from a direct comparison of cumulative and unfolding IRT methods as well as classical test theory (CTT) in terms of measurement precision, model fit, and construct and criterion validity by Chernyshenko et al. (2007).

The application of an unfolding model to real survey data is considered in the current investigation because many researchers have argued that responses to graded or dichotomous agree/disagree attitudinal items follow from an ideal point response process as opposed to a dominance response process. Thus, it is argued that the application of unfolding models to this type of data may be more appropriate than cumulative models because the assumptions of both unfolding models and data produced from ideal point response process are more congruent with each other (Andrich, 1988; Andrich, 1996; Andrich & Styles, 1998; Roberts & Laughlin, 1996; Roberts et al., 1998; 1999; 2000; van Schuur & Kiers, 1994). Roberts et al. (1998; 1999) detailed the consequences of applying a cumulative model to data of the unfolding type, where the model and data have competing underlying assumptions. Stark et al. (2006) also examined and tested which of the two models, cumulative or unfolding, best described personality data. Andrich (1988) described the inconsistency between data collection and analysis that many researchers use. For example, the predominant survey design method that is employed is the Likert scaling methodology (as opposed to a Thurstone approach), which implies use of a cumulative model. Observed data collected by asking respondents to disagree or agree with an item does not necessarily involve a direction, implying an underlying ideal point response process, thus suggesting the application of an unfolding model.

The body of research that exists within the context of unfolding IRT models focuses on improving the measurement of non-cognitive (preference, attitudinal, personality) traits by informing scale construction and development methodologies, and appropriate scoring and scaling procedures. Emphasis on improving measurement using models that have not been traditionally used comes from the fact that non-cognitive measurement has not been studied to the extent that cognitive measurement has, in the context of the ideal point response process and unfolding models. Non-cognitive measurement has been executed primarily through the use of cumulative IRT models and classical test theory models. Improvements in noncogitive measurement would necessarily improve the reliability of scores yielded by an instrument. The enhancement of non-cognitive measurement and scale construction would presumably result in an increase in accuracy and validity of decisions that are made based on scores. Improved measurement and better informed decisions based on scores have certain implications. Non-cognitive measurement has important roles in a variety of disciplines including many facets of psychology (i.e., developmental, industrial/organizational, social, and abnormal) and in education. Often times, results from questionnaires and surveys that measure personality and/or attitude are used to make educational policy decisions, develop treatment plans for the psychologically ill, or to make hiring decisions for prospective employees.

*Alternative Approaches for Assessing Model-Data Fit*

It has been recommended by Andrich (1996) and Chernyshenko et al. (2007) that comparison of the performance of cumulative and unfolding IRT models using simulated data and real data from a variety of disciplines would contribute to a greater understanding of the generalizability of unfolding models and to methodologies for improved measurement of non-cognitive constructs. Further, Chernyskenko et al. (2001) and Stark et al. (2006) argue for continued investigation into methods, alternative to the Likert methodology, for scale construction and into the consequences of model-data misfit. In all of the investigations mentioned in this literature review, either "truth" was known with the generation of simulated data, or prior parameter estimates existed for the same measure, providing a reference for comparison. Comparisons between the two types of IRT models (cumulative and unfolding) involve person and item parameters, in addition to graphical and analytical measures of model-data fit. The predominant statistical approach to assessing and comparing model fit includes the calculation of a chi-square distributed statistic. Item level chi-square distributed fit statistics are also frequently calculated as measures of both model assumptions and statistical fit within the context of cumulative and unfolding IRT models. Relative comparisons of model fit using chi-square statistics are not been possible, however, due to the fact that cumulative and unfolding models are not nested, thereby inhibiting the use of conventional and familiar statistics such as a log-likelihood ratio statistic.

The calculation of information theory-based measures (also referred to as information theory-based criteria or statistics) can provide an additional, unique source of evidence, in conjunction with other measures of fit, to assist in determining model selection. Such information theory-based criteria employ a penalty for complicated statistical models such that comparison of models with varying number of parameters is possible. A limitation of the widely used chi-square distributed statistics is such that models with more parameters to be estimated are more likely to fit better. These statistics are inconclusive in that the fit could truly be better, or simply an artifact of the number of parameters. Information criteria can be calculated to overcome this limitation of chi-square distributed statistics for non-nested model comparison where models vary in complexity.

When maximum likelihood methods for item parameter estimation are employed, it is possible to calculate criteria such as the AIC (Akaike, 1974) and the BIC (Schwarz, 1978). Bayes model selection criteria such as the Bayes Factor (Gelfand & Dey, 1994), the pseudo-Bayes Factor (Geisser & Eddy, 1979) or the deviance information criterion (DIC; as cited in Kang et al., 2005) are appropriately employed when methods for Bayes computation are used for parameter estimation. Perhaps the information theory-based criteria are most visible within the structural equation modeling literature, namely the AIC when examining predictive fit indices (Kline, 2005). They are considered predictive in that interpretation of fit is assessed within the context of how the model would fit data produced by repeated random samples drawn from the population as the observed sample (Kline, 2005). The AIC is also considered a parsimony-adjusted measure in that a penalty function for overparameterizing is incorporated in the equation. Akaike (1974) defined the criterion AIC of $\theta$ as:

$$\text{AIC}\ (\hat{\theta}) = (-2)\log\text{ (observed likelihood)} + 2k \tag{7}$$

"where *k* is the number of independently adjusted parameters to get $\hat{\theta}$" (Akaike, 1974, p. 719). An arguably more stringent measure of model fit is the BIC as sample size is directly considered where the calculation of the BIC criterion follows:

$$\text{BIC (model)} = (-2) \log \text{ (observed likelihood for a given model)} + p(log\text{N}) \qquad (8)$$

where N is the sample size. Early contributors to research for model identification and prediction unanimously emphasize the expression of a model, $\theta$, as a probability distribution and consider "fitting a model to the data as estimating the true probability distribution from the data and treat the estimation and the evaluation of a model together as one entity rather than separating them" (Bozdogan, 1987, p. 347).

Within the context of IRT measurement models, calculation of information theory-based statistics has predominantly been used for comparing the fit of latent class models (Bockenholt & Bockenholt, 1991; Houseman, Coull, & Betensky, 2006; Lin & Dayton, 1997) mixture IRT models (von Davier & Yamamoto, 2004) and penalized latent variable models (Haberman, 2006; Houseman, Marsit, Karagas, & Ryan, 2007). Interestingly, the utility of information theory-based statistics for comparing the fit of IRT models to assist in the determination of model selection in applied research appears largely within the health sciences literature, specifically for the application of latent class models to presumed high-dimensionality data with relatively small samples (Houseman et al., 2006; Houseman et al., 2007). Briefly, Houseman et al. (2006) proposed a penalized latent class model that circumvents the problem of "the number of conditional probabilities that can be considered without overfitting the data" (p. 1063) using latent class models. They proposed a method of deriving a constraint parameter; a parameter that is a direct function of the dimensionality of the data and the number of classes. Houseman et al. (2006) use both the AIC and BIC, among other information criteria, to estimate the number of latent classes, one parametric component, used in their penalized method. In a follow up study, Houseman et al. (2007) proposed a very similar methodology for the purpose of "data-driven model selection" (p. 1275), except the models were described as IRT models with the unobserved variables treated as continuous random, as opposed to categorical random. The AIC was again used to derive one component of the penalty function. Houseman et al. (2007) conducted a simulation study assess the functionality of the penalized likelihood models, applied their model and two Bayesian models (including the corresponding Bayes approach to their proposed method) to real data.

Finally, Hardouin and Mesbah (2004) extended the AIC to be used with multidimensional and non-parametric IRT models. They proposed a new multidimensional Rasch-type model, called the multidimensional marginally sufficient Rasch model (MMSRM), for the purpose of informed and improved Quality-of-Life scale construction over traditional factor analytic models. Their efforts also focused on the problem of unidimensional, parametric, IRT models exhibiting poor fit to data generated from measures like the Quality-of-Life scale. Implementing simulations to test their model, Hardouin and Mesbah (2004) used the AIC as a measure of fit between their proposed multidimensional IRT model and the Mokken scale procedure. They concluded that regardless of the model from which data were generated, their model generally performed better in correctly classifying items to respective subscales than the Mokken scale procedure (Harouin & Mesbah, 2004).

The use of information statistics for latent trait model selection within the context educational measurement using familiar IRT models such as the PCM, RSM, or GPCM (Kang & Cohen, 2007; Kang, Cohen, & Sung, 2005) is visible, though to a seemingly lesser degree (Ostini & Nering, 2006). Further, Takane (1996) employed the AIC in assessing the fit of his proposed multidimensional IRT proximity model for unordered categorical data; data for which unfolding IRT models would be most appropriately applied. The derivation and implementation of similar, though more statistically complex, penalty functions can be found in the educational measurement literature in the context of decision/classification accuracy with continuous predictor variables (Haberman, 2006). von Davier and Yamamoto, (2004) used information criteria (i.e., AIC and BIC) for the comparison of multi-group IRT models with partially missing data using their proposed model which "integrates multigroup IRT models and discrete mixture distribution IRT models into a common family of psychometric models" (p. 391).

Application of information theory-based criteria is conventional and appropriately used with a range of models (i.e., latent trait IRT, latent class IRT, structural equation models including confirmatory factor analysis). Although researchers from disciplines such as public health, educational, and psychological measurement commonly use information criteria as a tool for model selection, comparison, evaluation, and fit, such criteria are not without limitations. The AIC is appropriate for use with maximum likelihood estimates, though is criticized for not being "asymptotically consistent since sample size is not directly involved in its calculation" (Lin & Dayton, 1997, p. 251). The AIC criteria also tend to favor complex models when the sample size is large (Bozdogan, 1987). McDonald and Mok (1995) heavily criticize the AIC for use with both sufficiently large and small samples and claim that "the

AIC behaves just like the chi-square significance test itself" (p. 33). However, other information criteria that account for sample size such as the BIC (Schwarz, 1978) or Bozdogan's (1987) consistent Akaike information criterion (CAIC) have been derived to overcome this problem. Nonetheless, there is a general consensus that adhering to the principal of parsimony is necessary and that "there is no single criterion which will play the role of a panacea in model selection problems" (Bozdogan, 1987, p. 368).

### *Summary*

Although some advances have been made in the measurement of non-cognitive traits, more work is necessary as it relates to the application of unfolding models on a variety of data sets within a variety of contexts to better understand the processes that produce or govern observed data. Many scholars agree that research about non-cognitive measurement is not lacking, however, research that examines non-cognitive data within the context of unfolding models is sparse. Stark et al. (2006) provided a possible explanation for limited use of IRT models that assume an ideal point response process: "theoretical and computational complexity has impeded the development and application of ideal point methods" (p. 27). Following that, Coombs's (1951, 1964) original analysis for unfolding models was deterministic and "was laborious for more than four statements" (Andrich & Styles, 1998, p. 455). It is also generally agreed that the Likert methodology has been dominant over the Thurstone approach in non-cognitive measurement because the former does not require the labor-intensive step of scaling items. To date, parametric models for unfolding data have been developed, and computer software is available to analyze data using these measurement models.

Although data are usually collected through items that ask a respondent to agree or disagree, where often a direction is not inherently obvious, analysis of such data use cumulative IRT models that assume a dominance response process to govern the observed data (Andrich, 1988; 1996). For the purpose of increased measurement precision, accurate scores, valid interpretations of those scores, and appropriate and informed decisions based on scores, analysis of data presumed to be of the unfolding type must be analyzed using a procedure that makes parallel assumptions. Andrich (1996) argued that comparisons of cumulative models applied to Likert-type data to analyses using unfolding models “should prove instructive in improving the measurement of attitude and similar constructs” (p. 359).

Finally, there is general consensus among researchers that application of cumulative models to unfolding data can lead to the inaccurate measurement of people and that more focused efforts be made in non-cognitive measurement through the use of unfolding models including Andrich (1988), Andrich (1996), Andrich and Luo (1993), Andrich and Styles (1998), Chernyshenko et al. (2007), Luo, Andrich, and Styles (1998), Roberts & Laughlin (1996), Roberts et al. (1999; 2000), Roberts (2003), and Stark et al. (2006). This study addresses some of the omissions in the non-cognitive measurement literature and can function to fill this gap, especially given that real survey data are used--data upon which educational policy decisions have been and continue to be made for the purpose of improving teacher working and student learning conditions. Although the focus of this investigation is on the GGUM and the resulting parameter estimates and ICCs, comparisons will be made between analyses from the application of both cumulative and unfolding IRT models where the measurement of people and items can be examined across models. Additionally, the application of information theory-based criteria to three IRT and SEM models, in conjunction with other recommended statistical and graphical measures of model fit, will facilitate model comparisons.

# CHAPTER 3

# METHODS

This investigation used data from the 2006 administration of the North Carolina Teacher Working Conditions Survey (NCTWCS), an attitudinal measure of teacher perceptions of their working conditions. Eligible participants included all school-based licensed educators in the state of North Carolina including principals, assistant principals, teachers, and other teaching professionals like library media specialists or school psychologists.

## *Participants*

Although eligible participants for the NCTWCS included all school-based licensed educators, principals, assistant principals, and other education professionals (e.g., library media specialists, school counselors), for the purpose of this study, only responses from self-identified teachers were used because the questions are geared more towards teachers and teacher activities (i.e., teaching, preparing lesson plans) rather than principals, assistant principals, school counselors, or other school personnel. Many items on the NCTWCS are not applicable to respondents who are not teachers; therefore items could be omitted systematically by those respondents. Additionally, interpretation of the data would be difficult if self-identified principals, assistant principals, or school counselors responded to items that are targeted towards teaching activities. Finally, the decision to restrict the sample to teachers was made so that the integrity of the relationship between the underlying construct (i.e., teachers' perceptions of their working conditions) and the observed data is maintained.

The sample used for this study consisted of 65,031 self-identified teachers who responded to the NCTWCS, which represents 86% of the total sample of respondents (*n* = 75,615) that also included assistant principals, principals, and other educational professionals. The total number of unique schools in the sample was 2,365, of which 96 were charter schools and 19 were designated as special schools. The distinction between regular and special/charter schools is necessary because the response rates are vastly different for the two types of schools. The average response rate for the 115 special and charter schools was 19.27%, whereas the average response rate for the other 2,250 schools was 69.07%. Of the 2,250 regular schools, only 304 schools had a response rate of less than 40%.

Although not used in this investigation, the total number of principals who responded to the survey was 1392 of 75,615, or 1.84%. Assistant principals accounted for 1,544 (2.04%) of the total respondents and 7,449 (9.85%) of 75,615 self-identified as some other type of educational professional. It should be noted that a total of 115,105 people were eligible for the survey, and a total of 75,615 responded, for an overall response rate, across all categories of occupation of 65.69%.

By design, the NCTWCS was administered so that responses would be both confidential and anonymous. It is therefore not possible to link any response string to a particular respondent. The data set does contain a single identifier: the designated school code to which respondents were assigned at the time of data collection.

### *Measure*

In 1996, the North Carolina General Assembly established the North Carolina Professional Teaching Standards Commission. The responsibility of the Commission is to determine high standards for North Carolina teachers and the profession. The North Carolina Professional Teaching Standards Commission has written the Core Standards for All Teachers in North Carolina, Standards for Working Conditions in North Carolina Schools, and Professional Development Standards (North Carolina Professional Teaching Standards Commission, 2006).

In 2001, development of the NCTWCS began as a part of the governor's Teacher Working Conditions Initiative. The North Carolina Professional Teaching Standards Commission and the North Carolina State Board of Education conducted research and focus groups to develop 30 working condition standards for schools in the five overarching categories of time, empowerment, professional development, leadership, and facilities and resources. The categories were named or developed as a result of the focus groups that were conducted with more than 500 teachers. In 2002, an original survey, consisting of 29 items about working conditions within the five categories and in paper format, was distributed to every licensed public school educator in North Carolina.

After some revisions, the survey was administered in the Spring semester of the 2003-04 academic school year. The mode of administration changed to self-administered and web-based. Additionally, 33 working condition items were added for a total of 72 working condition items, plus eight demographic questions. The NCTWCS administered in 2006 comprised of a variety of items including: Likert-type with five response options (Strongly Disagree, Disagree, Neither Agree Nor Disagree, Agree, and Strongly Agree), Yes/No, Check All That Apply, and frequency items. Various items that elicit information about respondent demographics were included at the end of the survey. The mode of administration of the 2006 survey was the same as the 2004 administration. Anonymity was ensured by assigning all eligible participants a randomly generated access code that was required to begin the online survey. The scoring procedures for this particular survey are not publicly available; however, it appears that the scoring method was a simple summation of the item scores to yield a total scale score. It was also assumed that no reverse scoring was conducted because no negatively worded items appeared on the survey. A positive response (i.e., agree) to any Likert item would indicate a more positive attitude towards teacher working conditions. The 2006 NCTWCS can be found in Appendix A.

According to available documentation about the NCTWCS, the intent of the survey was to understand the factors that influence teachers' perceptions of their working conditions, as defined by the five domains, as previous research has shown that poor working conditions (i.e., lack of administrative support, lack of a collaborative atmosphere) contribute to teacher attrition. The assumption, then, is that improving working conditions may reduce teacher turnover, and thus improve student learning and achievement. The five domains of the NCTWCS were developed as a result of focus groups conducted with more than 500 NC teachers by the North Carolina Professional Teaching Standards Commissions. Although statistical measures of validity (i.e., criterion-referenced validity, convergent validity)  are not available, estimates of the reliability of the total scale (i.e., Likert items of the combined five domains) and of each of the sub-scales using Cronbach's coefficient alpha for both the raw and standardized item responses were calculated and are reported in Table 3.1.

Table 3.1

*Cronbach Coefficient Alpha for Likert items within the five domains (Total), the Empowerment and Leadership subscales*

| | | Cronbach's Coefficient Alpha | |
|---|---|---|---|
| | Number of items | Raw Variables | Standardized Variables |
| Scale | | | |
| Total | 52 | .966 | .966 |
| Empowerment | 13 | .897 | .897 |
| Leadership | 21 | .959 | .959 |

*Models*

As described in Chapter 2, a variety of unfolding IRT models exist for examining non-cognitive data. The GGUM was selected because it models subjective responses of agree and disagree (and variants of those responses) from above and below, because item discrimination and category threshold parameters are estimated, and because the probability of item endorsement ranges, theoretically, between zero and one. One of the limitations of Andrich and Luo's (1993) HCM is that the maximum probability of item endorsement is .5, even when the person and item location on the latent trait are coincident. In addition, GGUM was selected instead of Roberts and Laughlin's (1996) GUM because an assumption was made that the item response options are not interpreted and used equally by the respondents of the NCTWCS across all items and across all respondents.

The partial credit model was selected over other Rasch models for polytomous data such as Andrich's (1978) rating scale model because of the similarity of the PCM and the GPCM. The GPCM was chosen over other non-Rasch models for polytomously-scored data such as Samejima's graded response model (Samejima, 1969) or Bock's (1972) nominal response model because of the association between the GPCM and the generalized graded unfolding model. Although any cumulative IRT model appropriate for polytomous data could be used to estimate the subjective response categories (SRCs) in the GGUM, Muraki's (1992) generalized partial credit (GPCM) model was used to parameterize these functions in this investigation. To maintain consistency as much as possible, the GPCM was used both as the model to measure the SRCs within the GGUM, and as the cumulative model for comparison with the PCM and GGUM.

The decision to implement a fourth scaling method was made so that comparisons of the IRT analyses could be made to the possible current scoring procedure.  A total raw score scoring method can inhibit score comparisons due to the lack of equal interval data and a meaningful scale. The items in the NCTWCS are Likert-type which do not necessarily yield interval data (i.e., the difference between Disagree and Strongly Disagree is not necessarily the same as the difference between Agree and Strongly Agree). Additionally, the use of raw scores precludes determining an individual's standing on the latent trait when the individual earns an extreme (i.e., maximum or minimum possible) score on the survey. Finally, measurement precision for extremely scoring respondents is low when using raw scores. Thus, the fourth scaling method used was a CFA approach. CFA is a more psychometrically sound approach because the raw, ordinal data are transformed and placed on a more meaningful scale. Specifically, the products of factor scores that resulted from the CFA were used to weight item responses, and then summed across the items for each person within each scale (i.e., Empowerment and Leadership).

## Data Analysis

Information about analysis procedures as it relates to each research question is provided in this section. Generally, this study examined the functionality of two cumulative and one unfolding IRT model, and where appropriate, comparisons to a fourth model were made. Although there are multiple sections of the NCTWCS that include various item-types, only 34 of the 52 Likert-type items within two of the five components of working conditions were considered in the analyses due to concerns of multidimensionality. Specifically, the 13 items that measure Empowerment and the 21 items that measure Leadership were used. The two components were selected based on the substantively most interesting and important factors to both teachers and to the policy makers who implement changes within schools based on the survey data. An example item from the Empowerment subscale reads: "Teachers are centrally involved in decision making about educational issues." An example item from the Leadership scale reads: "The school leadership consistently enforces rules for student conduct." Items measuring Time, Professional Development, and Facilities and Resources were not used in this investigation. All items in the selected subscales required respondents to select a response option to indicate strength or level of agreement choosing from the five response options: Strongly Disagree, Disagree, Neither Disagree Nor Agree, Agree, and Strongly Agree. All analyses were conducted on each scale (i.e., Empowerment and Leadership) separately, resulting in two sets of analyses.

Although the total number of respondents was 65,031, the original single sample could not be used for analyses due to sample size constraints with the GGUM2004 software (maximum sample is 2,000). As a result, 10 simple random samples of size 2,000, selected without replacement from the original sample, were constructed and used in all analyses. The 10 samples were then separated by scale (i.e., Leadership or Empowerment). The sample selection was executed using the SURVEYSELECT procedure and SRS method in the statistical software package SAS 9.1 (SAS Institute, 2002).

### *Estimation Procedures*

All IRT analyses required the use of statistical software to estimate IRT item and person parameters. The program PARSCALE 4 (Muraki & Bock, 1997) was used for application of the partial credit and generalized partial credit models. The program GGUM2004 (Roberts et al., 2000; Roberts et al., 2006) was used for the application of the generalized graded unfolding models. Other software programs exist for cumulative IRT calibration, however, PARSCALE was used for both the PCM and GPCM to facilitate comparisons across models.

With regard to the algorithms used in the IRT estimations, briefly, the marginal maximum likelihood method (MML; Bock & Aitkin, 1981) is implemented in PARSCALE for item parameter estimation. The Expectation-Maximization (EM; Dempster, Laird, & Rubin, 1977) algorithm is integrated in the derivation of maximum likelihood solutions. Either maximum likelihood or Bayes estimation is feasible with PARSCALE for estimating person parameters. The *expected a posteriori* (EAP) method was used in this investigation for estimating the person parameters, $\theta$. A type of prior distribution is necessary for the estimation of $\theta$; a normal distribution was specified for this investigation. Additional information on estimation in PARSCALE can be found in du Toit (2003).

The GGUM2004 software also estimates item parameters using a marginal maximum likelihood method. According to the GGUM2004 technical manual, “the solution algorithm parallels Muraki's (1992) procedure used in the generalized rating scale model and is based on an expectation-maximization (EM) strategy" (Roberts & Shim, 2008, p. 5). In order to estimate $\theta$, specification of a prior distribution is necessary in any analysis where $\theta$ is treated as a random variable. For both the GGUM2004 and PARSCALE analyses, a standard normal prior distribution was assumed for $\theta$. The EAP method for $\theta$ estimation was also used in the GGUM analyses. Further details about the software are available in the *GGUM2004 Technical Reference Manual* (Roberts & Shim, 2008). Finally, the software used to conduct the CFAs and calculate factor scores for each sample was LISREL 8.8 (Joreskog & Sorbom, 2006). Additional analyses including the calculation of chi-square statistics for all possible combinations of item pairs and item triplets, and principal components analyses, both necessary for testing model assumptions were conducted using SAS 9.1 (SAS Institute, 2002).

### *Omitted Data*

Omitted item responses for all four models (GGUM, PCM, GPCM, and CFA), were treated in a similar manner. In order to determine how to proceed with analyses and the treatment of missing data, it is necessary to establish the extent and to assess the randomness of missing data. Missing responses were examined at the item level, within each sample across both scales, for systematically missing responses. Within the 10 Empowerment samples, the two most frequently omitted items read: "In this school we take steps to solve problems" and when prompted to rate how large a role teachers have in: "Devising teaching techniques." The first of the two items yielded a range of 22 to 36 missing responses across the 10 samples, representing 1.1 % to 1.8% of a sample omitting this item, and 20 to 37 people omitting the second item, resulting in 1.0% to 1.85% missing data for that item..

Three approaches were taken to assess for patterns of systematic omissions within the 10 Empowerment samples. First, those respondents missing both items were separated from those not missing both items across the 10 samples. Group means could not be compared due to the extremely low number of individuals across the 10 samples omitting both items (*n* ranged from one to three respondents). Second, those respondents omitting the first item were separated from those not omitting that item. The same procedure was followed for the second item. Again group means could not be compared due to the low samples for those omitting either the first or second item. For example, in the first Empowerment sample, 24 people omitted the first item and 1,976 did not. In the first Empowerment sample, 31 people omitted the second item and 1,969 did not. Group comparisons were not made due to the disparity of sample sizes, as the smaller sample could potentially introduce bias into the estimates, and suppress power to detect true differences. Finally, to assess for an inverse relationship of omissions between the two items, each of the 10 Empowerment samples were subset into two groups: those omitting the first item, and not the second; then sub-setting the sample into those who did not omit the first item but did omit the second. Again, the sample sizes were vastly different preventing group mean comparisons, in that there were very few (approximately 25 missing item responses across the 10 samples) in the sample that omitted an item. As a result of the above considerations, it was concluded that no systematic pattern of omissions existed in the Empowerment samples.

Only one item, “The school leadership makes a sustained effort to address teacher concerns about facilities and resources”, was consistently omitted across nine of the 10 Leadership samples at a rate of between 33 to 45 people in a given sample, representing 1.65% to 2.25% of the samples, respectively. As a result of the relatively low percentage that the missing data accounted for in both the Empowerment and Leadership samples, these omissions were not considered problematic and the decision was made that missing data were not systematic in this investigation.

According to the *GGUM2004 Technical Reference Manual*, “the GGUM2004 software accommodates missing item responses by treating missing data as missing at random given θ. In the context of the GGUM family of models, this means that the any missing item responses are simply ignored when calculating the likelihood of a given response vector (e.g., either *Ls(Vf )* or *Lj(Vf)* in the preceding equations)” (Roberts & Shim, 2008, p. 19). According to the PARSCALE manual, “omitted responses are treated as not-presented” (DuToit, 2003, p. 336).

### *Analysis Overview*

For all items, item characteristic curves (ICCs) were compared across the three IRT models. The ordering of item location parameters was examined across the cumulative and unfolding models, as it has been argued by Chernyshenko et al. (2007) that the item location parameters yield information about item content within the context of unfolding models, a result that does not hold for cumulative IRT analyses. Additionally, special attention was given to the shape of the ICCs. Specifically, lack of monotonicity for the cumulative IRT results would suggest a disparity between the model and data. The opposite is not true, however, for unfolding models such that if monotonic ICCs are found, the interpretation that an unfolding model is inappropriate would not necessarily be true. The GGUM software used in this investigation is flexible enough to model items with monotonic ICCs, even when an ideal point response process is responsible for the observed monotonic ICC. Non-monotonicity in the ICCs resulting from application of a cumulative model would warrant reconsideration about how data should be treated and scored. Finally, for person estimates, the general locations of person estimates and the ordering of those estimates were examined across the three IRT and the CFA models, with attention to those respondents with extreme estimates in either direction.

*Research Question 1*

The first research question related to examining the location of the items on the underlying latent traits (i.e., teacher perception of empowerment and leadership) across three IRT scaling methods (partial credit model, generalized partial credit model, and the generalized graded unfolding model). Item parameters, especially the location, or *b* parameter, and the ordering of those parameters were compared across methods. Graphical representations of item parameters were prepared, along with correlations of the ordering of items across models. Any gaps where no items existed in a particular region or interval on the latent scale would indicate a lack of measurement precision in those regions.

Examination of item location can yield important information for future survey development and design. If the items on the attitudinal measure were constructed using a Likert methodology and modeled using an unfolding model assuming an ideal point response process when individuals respond to items, then items will generally be located at the ends of the latent trait continuum. Within the Likert methodology for test development, criteria for maintaining items on a scale include high point biserial correlations, high factor loadings, and monotonically increasing ICCs. Generally, items that measure more of a neutral attitude generally do not meet these criteria. Items meeting the criteria tend to be worded in more extreme terms (i.e., items that express both extreme positive and extreme negative sentiment with respect to the latent trait). The scoring of Likert-type items requires the reverse scoring of items that reflect a negative attitude.

Close examination of item parameters and the relative ordering of the items has been recommended by Andrich (1995a) who has observed that the theoretical ordering of items "is particularly important when the model, which reflects the response process, is single-peaked because there are always two person locations that give the same probability of a positive response" (p. 275). If the items on the attitudinal measure were not constructed using a Likert methodology and more neutral items existed on the survey, then item locations on the latent trait will be more similar to each other across all three scoring and scaling methods, than if a strict Likert methodology were used. Specifically, item locations will generally be more centrally located, or at least more dispersed across the attitude continuum as opposed to located towards the extreme values of the latent trait.

*Research Question 2*

The second research question concerned examination of person estimates on each of the two scales (i.e., location or theta IRT parameters in the GGUM, PCM, and GPCM analyses, and the composite score in the CFA analyses) on the underlying latent traits across the four methods of scaling. Graphical representations were developed to visually examine the relationship between the estimated theta distributions across models. Examinations of the theta distribution across models were facilitated with the use of scatterplots by plotting the person estimates (i.e., thetas) for the GGUM, and those resulting from each of the GPCM, the PCM and the CFA analyses. Because the theta scales are different across the unfolding IRT, cumulative IRT and CFA models, simple scatterplots and the nonparametric correlation for ranked data, Kendall's Tau, were derived. Finally, for each scaling method, the continuous theta scale was transformed into a discrete scale with the use of quintiles. This allowed for examination of the person distribution at specified intervals. Using the five theta categories, the calculation of 5 X 5 cross tabulation tables under the combination of GGUM with the other three models across all samples (n = 10) within each scale (n = 2) facilitated examination of the joint theta distributions. Close attention was given to those respondents estimated to be located at moderately and very extreme locations on the latent trait when comparing across scaling methods. Examination at those levels of the latent trait is necessary because researchers (Roberts et al., 1998, 1999; Stark et al., 2006) have shown that the greatest disparity between cumulative and unfolding models occurs within these regions of the latent trait. Therefore, if the assumption that the observed data followed from an ideal point response process was true, it would be hypothesized that a discrepancy would exist between the cumulative and unfolding models only for respondents with extreme responses to moderately positive and moderately negative items.

### *Research Question 3*

The third research question required an examination of the shape and location of the ICCs to assist in the determination of whether an ideal point or dominance response process operated in these data. Again, the ends of the ICCs from all analyses using the IRT models were examined closely for discrepancies. Specifically evidence of non-monotonicity of the ICCs from the cumulative analyses and evidence of monotonicity of the ICCs from the unfolding analyses is especially important in determining if and how outcomes and results would differ across the different scaling methods.

It was hypothesized that item characteristic curves (ICCs) for the dominance IRT models would be monotonic increasing functions. Further, ICCs associated with dominance models would exhibit non-monotonicity for items containing neutral or non-extreme sentiments towards teachers' perceptions of empowerment and leadership in their school. Because most of the 13 items on the Empowerment scale and 21 items on the Leadership scale are seemingly neutrally worded, it was presumed that the majority of the ICCs resulting from the application of dominance IRT models would exhibit non-monotonicity. With respect to the unfolding model outcomes, it was presumed that the ICCs would appear non-monotonic, and relatively single-peaked for items of neutral and moderately positive and negative sentiment about perceptions' of teacher empowerment and leadership. ICCs associated with unfolding models were hypothesized to exhibit monotonicity only if at least the large majority of all respondents were located to one side of the item on the latent trait (i.e., extremely homogeneous sample). Finally, ICCs for both cumulative and unfolding models should appear monotonic only for extreme responses to items that contain an extreme sentiment. However, none of the items used in the current study would be categorized as extremely positive or extremely negatively worded. Consequently, monotonic ICCs were not expected for both types of IRT models for a given item.

### *Research Question 4*

The fourth research question had two parts that pertained to: 1) testing model assumptions within the domain of both cumulative and unfolding models; and 2) statistical model fit for the generalized graded unfolding model, relative to the other scaling methods. With regard to model assumptions, statistical procedures for summarizing patterns of correlations among item responses (e.g., structural equation modeling techniques (SEM), principal components analysis (PCA)) are commonly employed for assessing test dimensionality within the context of both cumulative and unfolding models. Model specification differs, however, across the two types of IRT models. A single latent factor can be expected to explain a set of observed item responses if data follow from a dominance response process and two latent factors resulting from a principal components analysis (PCA) should result if data follow from an ideal point response process. Based on the procedures of Davison (1977), Nandakumar, Hotchkiss, and Roberts (2002) and van Schuur and Kiers (1994), linear factor analytic methods (i.e., PCA) were applied and pattern coefficients of items were examined to assess unfolding dimensions. Plots of pattern coefficients were constructed as a visual aid to assist in determining the structure of the data. According to Davison (1977) and Nandakumar et al. (2002), data are presumed to be of the unfolding type when a PCA of inter-item correlations yields a two-factor structure and, when plotted, pattern loadings for those two components form a semi-circle. Additionally, root mean square residuals for each item were calculated to allow for relative comparisons across items, where smaller values represent better fit. This process was undertaken to examine the fit of the model at the item level and to examine the property of local independence.

Pertaining to the component of the fourth research question that addresses the fit of GGUM relative to the other three scaling methods, information theory-based statistics were calculated as opposed to the commonly used chi-square distributed statistic, because the latter cannot be used for relative non-nested model comparisons. Currently, no chi-square distributed fit statistic (i.e., log-likelihood ratio) exists that allows for relative comparisons of model-data fit for non-nested cumulative and unfolding IRT models, such as for the three IRT models used in this investigation. The PCM and GPCM can be compared using a chi-square fit statistic using the difference in chi-square and degrees of freedom values because they are nested models. However, such a comparison is not warranted and is tangential to this investigation as the focus is on the monotonicity or non-monotonicity of the ICCs.

The information theory-based fit indices calculated in this investigation were the AIC and BIC. Both of these indices consider the non-nested characteristics of the models and the additional parameters in the more complicated models (i.e., GGUM, GPCM). Additionally the BIC directly considers sample size and tends to favor simpler models than the AIC. Better fit using these statistics is indicated by smaller AIC and BIC values.

## Summary and Limitations

The analyses conducted in this investigation focused on both a relatively new method, and more familiar IRT approaches for analyzing polytomous attitudinal data. Although cumulative IRT models have been used extensively in analyzing non-cognitive, polytomous data, unfolding IRT models warrant attention because of their potential for improving scale construction and score interpretation. As a result of recent advances in the derivation of probabilistic models, and software capabilities, applied research using unfolding IRT models is in its relatively early stages, compared to cumulative models. One purpose of this investigation was to contribute to the methodological research surrounding the relatively novel approach to measurement using unfolding IRT models; models which could prove useful and informative from both psychometric and practical perspectives.

The appropriateness of the application of several parametric, unidimensional IRT models to real survey data from the administration of the NCTWCS was investigated by considering several guiding questions. The first two research questions dealt with the location of the respondents and the items on the same, unidimensional latent trait scale. Two components of the original scale were treated as two distinct latent traits. This approach allowed for exploration of the functioning of all models across two scales while partially controlling for the confounding effects of multidimensionality. Additionally, 10 simple random samples were selected from the original sample and all analyses were performed on all samples across the two scales. This allowed for the display of sampling distributions of outcomes. Other characteristics of the items were examined such as the shape of the ICCs and category probability plots across the three IRT models. The distribution of person parameters was also examined by making the continuous theta distribution discrete using quintiles. The joint distribution of theta under the GGUM paired with the other scaling methods was facilitated using cross tabulations. Finally, model assumptions were tested and relative fit across the four models were compared using AIC and BIC statistics.

Results from these analyses could help to inform future versions of NCTWCS. A close examination of item discrimination, item location on the latent trait (which can be interpreted as a measure of intensity of item content), and item fit were made. Because the purpose of the NCTWCS is to ascertain teachers' perspectives about their working environment, efficiency is achieved with the least amount of items that measure the entire spectrum of the latent trait. It is usually the case in non-cognitive measurement that measurement is necessary across the latent trait in its entirety, as opposed to cognitive measurement where, generally, precision and item (and test) information are often focused on an interval(s) of the latent trait, usually around one or more cut scores. Careful survey construction could increase the efficiency and efficacy of the measurement of the latent trait.

Within the context of test development in personality measurement, Stark et al. (2006) note that “inclusion of just a few items that do not meet the assumptions of dominance models can markedly change the rank order of high-scoring individuals and, thus, potentially undermine the utility of personality measures in applied settings” (p. 37-38). The examination and evaluation of item and person location on the latent trait in this investigation would also inform the survey development process and direct attention to areas on the scale that require more precise measurement.

The methodology of this investigation was limited by several factors. First, the sample in its entirety, (N =  65,031) could not be used due to software constraints. The reduction of sample size also prohibited parameter estimates for all persons in the sample. One advantage of previous research studies by Andrich (1988), Andrich and Luo (1993), Habing et al., (2005), Hoijtink (1991) and Roberts et al. (2002) is that they had previous parameter estimates for particular scales allowing for direct comparisons between IRT models and parameter estimates. In this investigation, however, parameter estimates did not exist for the NCTWCS, prohibiting any relative comparisons and absolute decisions to be made with GGUM parameter estimates.

With regard to the measures, all items have five response options, including the middle category, Neither Agree nor Disagree. According to researchers including Andrich and Styles (1998), the middle category does not necessarily function as the mid-point between two adjacent response options. According to Andrich (1996) the middle category--intended within the Likert methodology of item writing to function as an undecided/ambivalent response option--has “consistently posed problems” (p. 362) in interpreting the meaning of an ambivalent response to a seemingly extreme item. In the current investigation there was no meaningful way to collapse responses and as result all five response categories were retained.

Despite these limiting factors, this investigation made use of a relatively new IRT model for interpreting real attitudinal data within the context of educational policy, employed criteria for determining relative model fit, closely examined the estimated latent trait distributions resulting from application of the four models, across all simple random samples and within scales. Overall, this investigation addressed recommendations previously made by researchers such as: the use of real data when examining the functioning and appropriateness of unfolding IRT models, the implementation and comparison of both cumulative and unfolding IRT models, and the application of unfolding IRT models within different contexts, in this case educational policy. Comparisons were further supported with the use of two scales (i.e., Leadership and Empowerment). Although cited as a limitation, the random sampling involved in this methodology allowed for distributions of outcomes to be constructed and displayed across the 10 samples for each of the two scales. The analyses involved in this study allow for recommendations of modifications to survey construction which could presumably shorten the NCTWC survey and increase measurement efficiency Finally, not only are the implications immediate to the survey development procedures for the NCTWCS, but this investigation could have implications for assessing the knowledge, skills, and behaviors, necessary to successfully perform a school leadership position (i.e., principal) and to identify the critical tasks of a principal as part of a practice analysis.

# CHAPTER 4

# RESULTS

One unfolding model, two cumulative IRT models, and one structural equation measurement model were applied to attitudinal data from two subscales of the North Carolina Teacher Working Conditions Survey (NCTWCS). Model assumptions were tested and resulting parameter estimates were examined and compared using correlational analyses, chi-square fit statistics, information theory-based fit statistics, with a focus on the behavior and functioning of the relatively new generalized graded unfolding model (GGUM, Roberts et al., 2000). Ten simple random samples of size 2,000 were selected from the full sample of 65,031 self-identified teachers who completed the (NCTWCS) during the 2005-2006 academic year.

As with any parametric, probabilistic measurement model, certain model assumptions should be met in order to increase confidence in the accuracy of resulting parameter estimates. Testing of these assumptions is always a first step in any statistical analysis. Here, three different types of models were used: confirmatory factor analysis, cumulative IRT models, and an unfolding IRT model, all of which rest on assumptions. In the following sections, results will be presented and discussed for the Empowerment scale first; the second half of this chapter contains all analyses and results related to the Leadership scale.

## Empowerment Scale

### *Testing Dimensionality Assumptions for Cumulative Models*

Although presented first, the fourth research question investigated had to do with model assumptions and model fit. In practice, any investigation should begin with testing models assumptions. To test model assumptions, various procedures were conducted for all four scaling methods and are reported in this section for the Empowerment data. Within the context of CFA, assumptions include a linear relationship between observed variables (i.e., item responses) and unobserved factors (i.e., latent construct, or theta). Two requirements of CFA analyses include model identification and specification. Model identification ensures unique parameter estimates for all free model parameters. Two basic requirements must be met for a model to be identified: the number of moments must exceed the number of free parameters, and each latent factor must be assigned a scale (Kline, 2005). Model specification is more theoretical and deals with the direction of association between variables. In the current investigation, a one-factor model was specified using the 13 items that measured the construct, teachers' perceptions of teacher empowerment in their school. For model identification purposes, the factor loading was fixed at 1.0 in the analyses for the item that possessed the largest measure of variation. In the empowerment analysis that item was "Teachers are centrally involved in decision making about educational issues."

Numerous model-data fit indices (predictive, parsimony-adjusted, incremental) are calculated for each analysis including, but not limited to the root mean square error of approximation (RMSEA), root mean square residual (RMR), goodness of fit index (GFI), comparative fit index (CFI), Akaike information criterion (AIC), expected cross-validation index (ECVI), normed fit index (NFI) and the non-normed fit index (NNFI).

Model fit results from the single factor confirmatory analysis using the Empowerment items are reported in Table 4.1. As shown in the Table, the statistically significant (*p* < .05) model chi-square statistic and the RMSEA indicate that a single factor model does not fit well. The RMR, SRMR, and GFI also indices indicate less than adequate fit. The only indication of moderately good model fit is the GFI index. Finally, the matrix of inter-item correlations for the 13 items is presented in Table 4.2. Multiple indices, such as those reported, must be considered simultaneously in determining model fit, and for the Empowerment data, a single factor model with 13 items does not appear to fit at all well. Several methods exist for assessing test dimensionality within the context of cumulative IRT models. One common approach is the application of factor analytic methods. The results presented indicate that the assumption of unidimensionality within the context of the cumulative IRT models (CFA, PCM, and GPCM) may be violated to some extent.

Table 4.1

*Fit Indices for the One Factor Empowerment Model: Full Sample (n = 65,031)*

| Model $\chi^2$ | df | $\chi^2$/df | RMSEA | RMR | SRMR | NFI | GFI | Model AIC |
|---|---|---|---|---|---|---|---|---|
| 68348.838* | 65 | 1051.52 | .148 | .075 | .075 | .914 | .820 | 85394.63 |

*Notes:* RMSEA = Root Mean Square Error of Approximation; RMR = Root Mean Square Residual; SRMR = Standardized Root Mean Square Residual, NFI = Normed Fit Index; GFI = Goodness of Fit Index, Model AIC = Akaike Information Criterion
* p < .05

*Local Independence*

Related to the unidimensionality assumption, cumulative IRT models pose an associated assumption: local item independence. To examine this assumption, root mean square residual at the item level was calculated. This was tested on the entire sample for the PCM and GPCM models on both the Empowerment and Leadership samples. Smaller residual values are interpreted as better relative fit.

Table 4.2

*Item Level Residuals from PCM and GPCM Models: Empowerment Scale*

| Item | PCM | GPCM | Item | PCM | GPCM |
|---|---|---|---|---|---|
| 1 | 10.460 | 5.279 | 8 | 6.114 | 3.914 |
| 2 | 8.156 | 40.762 | 9 | 7.005 | 5.209 |
| 3 | 9.722 | 5.389 | 10 | 2.846 | 4.048 |
| 4 | 8.466 | 5.557 | 11 | 5.676 | 5.663 |
| 5 | 5.916 | 6.141 | 12 | 2.561 | 4.280 |
| 6 | 5.822 | 6.548 | 13 | 6.950 | 5.712 |
| 7 | 4.910 | 5.585 | | | |

Table 4.2 displays the root mean square residuals at the item level from both the PCM and GPCM models. Because of the sample size restriction and other software restrictions, chi-square item level likelihood-ratio fit statistics were used as a proxy for a measure of local independence for the GGUM. These chi-square distributed statistics were calculated for measures of model fit and are reported and interpreted in Table 4.15 in the Item Parameter Estimates section for the GGUM analyses, by sample, on the Empowerment data. To summarize the results shown in Table 4.3, items 1, 3, 4, 8, and 13 fit statistically well in most of the GGUM analyses, although these statistics and associated *p*-values, must be interpreted with caution. At the scale level, the GGUM did not fit statistically well in any analysis.

Table 4.3

*Pearson Product Moment Inter-item Correlations between the 13 Empowerment Items*

| Item | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | - | | | | | | | | | | | | |
| 2 | .674 | - | | | | | | | | | | | |
| 3 | .697 | .603 | - | | | | | | | | | | |
| 4 | .628 | .567 | .766 | - | | | | | | | | | |
| 5 | .402 | .365 | .405 | .392 | - | | | | | | | | |
| 6 | .430 | .429 | .378 | .361 | .273 | - | | | | | | | |
| 7 | .392 | .488 | .358 | .361 | .242 | .587 | - | | | | | | |
| 8 | .306 | .368 | .270 | .267 | .201 | .439 | .561 | - | | | | | |
| 9 | .464 | .393 | .446 | .412 | .330 | .434 | .371 | .365 | - | | | | |
| 10 | .333 | .265 | .303 | .286 | .221 | .298 | .224 | .192 | .401 | - | | | |
| 11 | .479 | .408 | .470 | .469 | .312 | .367 | .348 | .301 | .465 | .427 | - | | |
| 12 | .419 | .317 | .392 | .363 | .261 | .361 | .267 | .220 | .451 | .482 | .476 | - | |
| 13 | .498 | .402 | .489 | .462 | .305 | .430 | .365 | .274 | .472 | .374 | .487 | .517 | - |
| Item<br>Mean | 3.20 | 3.64 | 3.33 | 3.51 | 3.33 | 3.43 | 3.70 | 3.40 | 2.61 | 1.80 | 2.60 | 2.02 | 3.12 |
| S.D. | 1.13 | 1.07 | 1.10 | 1.05 | 1.08 | 1.00 | .99 | 1.11 | 1.06 | .97 | 1.10 | 1.00 | 1.04 |
| Var. | 1.27 | 1.15 | 1.22 | 1.11 | 1.17 | .99 | .98 | 1.23 | 1.11 | .94 | 1.21 | 1.00 | 1.08 |

Investigation of model assumptions, and to some extent, model fit, revealed that the CFA and cumulative IRT models exhibited moderately good fit, and the assumption of unidimensionality was reasonably met. Although absolute interpretations cannot be made using the RMSEA statistics, they revealed, at the item level, that the PCM fit better for half of the items, and the GPCM fit better (than the PCM) for the other half of the Empowerment items.

### *Unidimensionality under Unfolding Models*

Assessing the dimensionality of the data from the ideal point response perspective, a principal components analysis was conducted on the entire sample (*n* = 65,031) to assess dimensionality. The methods used to examine the dimensionality of each data set (Leadership and Empowerment) included examination of eigenvalues, final communality estimates, pattern coefficients, and plots of pattern coefficients resulting the application of principal components analyses with two components. Assessment and determination of dimensionality structure within the context of unfolding models is similar to determination within the context of cumulative models in that consideration of a variety of measures is necessary, where there are established criteria for some measures, and general heuristics for others. Generally, if an item level communality, generated from the first two components, is $\geq$ .3, then that item is not likely violating the assumption of unidimensionality (Roberts et al., 2000). Item level final communality estimates derived from a two factor component model for the Empowerment sample are reported in Table 4.4. Communalities for all items comprising the Empowerment scale were greater than .3.

The first two eigenvalues of the PCA should be larger than the remaining eigenvalues when the data are unidimensional, of the unfolding type. This criterion is a rule-of-thumb; no formal criterion exists to strictly quantify “large.” The first five eigenvalues from the Empowerment analysis, in descending order were: 5.889, 1.254, 1.163, .732, and .604.

Table 4.4

*Final Communality Estimates for the Empowerment items (i = 13)*

| Item | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | .654 | .550 | .667 | .614 | .308 | .614 | .756 | .683 | .478 | .340 | .507 | .459 | .514 |

The second eigenvalue was not substantially larger than the third eigenvalue. This is evidence that the assumption of unidimensionality within the context of cumulative IRT models is supported, as opposed to unidimensionality within the context of unfolding models.

A third procedure for examining dimensionality within the context of unfolding models is to plot the pattern coefficients that result from a PCA with two factors (i.e., components). A semi-circular pattern (i.e., simplex pattern) of the coefficients in a two-factor space is evidence that two linear principal components explain the pattern of data. Within the context of unfolding IRT models, there are two linear principal components for each unfolding dimension. The pattern coefficients from the two component PCA with the Empowerment samples are reported in Table 4.5.  Figure 4.1 displays the plot of pattern coefficients for the Empowerment scale. A distinct semi-circular pattern is not evident for the 13 Empowerment items, suggesting that two linear components are not responsible for producing the observed pattern of responses. All of the procedures for assessing unidimensionality within the context of unfolding data reveal that the responses do not unfold, and that application of unfolding IRT models may be unnecessary.

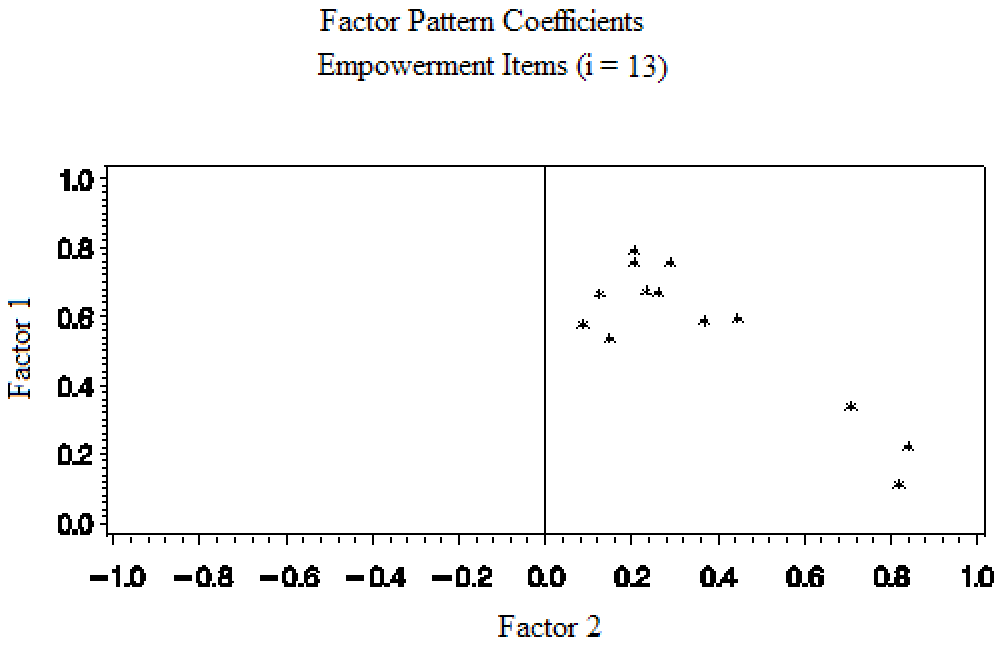
Table 4.5

*Factor Pattern Coefficients Derived from Two Principal Components: Empowerment*

| Item | Factor1 | Factor2 | Item | Factor1 | Factor2 |
|---|---|---|---|---|---|
| 1 | .755 | .289 | 8 | .112 | .819 |
| 2 | .593 | .445 | 9 | .586 | .368 |
| 3 | .790 | .206 | 10 | .577 | .088 |
| 4 | .756 | .205 | 11 | .672 | .235 |
| 5 | .534 | .149 | 12 | .665 | .126 |
| 6 | .338 | .707 | 13 | .668 | .262 |
| 7 | .221 | .841 | | | |

Figure 4.1

*Plot of Factor Pattern Coefficients for the Empowerment Scale*



## Development of the CFA Scale

The single factor CFA analyses were implemented as scaling methods and were selected to mirror the scaling methodology frequently used in current survey research. According to documentation in reference to the NCTWCS development and data manipulation, (Center for Teaching Quality, 2006), results from factor analyses were used to create domain averages across the sections, including but not limited to those of Empowerment and Leadership. To maintain a close resemblance to current scoring/scaling methodology, an assumption was made that a single latent trait, Empowerment, was measured by the 13 empowerment items. In the empowerment analysis, the item that consistently yielded the highest standard deviation reads: “Teachers are centrally involved in decision making about educational issues”, with the standard deviation ranging from 1.114 to

1.147 across the 10 empowerment samples. As a result, the factor loading of this item onto the Empowerment factor was fixed at 1.0 for all 10 Empowerment analyses.

The measures of the latent trait were transformed into observed variables by summing the products of the factor scores and item responses per respondent. The summation across the 13 Empowerment items functioned as the CFA Empowerment theta or measure of Empowerment trait per respondent. Various measures of model/data fit for the 10 empowerment samples are shown in Table 4.6.

Table 4.6

*Fit Indices for the One Factor Empowerment Model by Sample (n = 10 samples)*

| Sample | Model $\chi^2$ | df | $\chi^2$/df | RMSEA | RMR | SRMR | NFI | GFI | Model AIC |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 2152.176 | 65 | 33.110 | .146 | .075 | .075 | .915 | .819 | 2690.97 |
| 2 | 2248.426 | 65 | 34.591 | .152 | .078 | .078 | .912 | .808 | 2889.7 |
| 3 | 2096.821 | 65 | 32.259 | .145 | .077 | .077 | .907 | .821 | 2643.81 |
| 4 | 2188.376 | 65 | 33.667 | .147 | .075 | .075 | .911 | .817 | 2723.15 |
| 5 | 2182.586 | 65 | 33.578 | .147 | .075 | .075 | .914 | .817 | 2721.22 |
| 6 | 2325.526 | 65 | 35.777 | .155 | .080 | .080 | .900 | .801 | 3024.7 |
| 7 | 2568.795 | 65 | 39.520 | .144 | .073 | .073 | .917 | .824 | 2620.8 |
| 8 | 2026.286 | 65 | 31.174 | .145 | .075 | .075 | .916 | .822 | 2603.62 |
| 9 | 2054.238 | 65 | 31.604 | .145 | .073 | .073 | .916 | .822 | 2592.43 |
| 10 | 2233.273 | 65 | 34.358 | .149 | .077 | .077 | .913 | .813 | 2828.364 |

*Notes:* RMSEA = Root Mean Square Error of Approximation; RMR = Root Mean Square Residual; SRMR = Standardized Root Mean Square Residual, NFI = Normed Fit Index; GFI = Goodness of Fit Index, Model AIC = Akaike Information Criterion

The preceding results indicate that a single factor model did not fit the data well; therefore the assumption of unidimensionality within the context of cumulative IRT models was not satisfied. Further, it can be concluded that the 13 items do not measure the construct, Empowerment in the CFA model well either. Unidimensionality within the context of unfolding IRT models was also not met given that two components did not explain the data, and that the plots of the pattern coefficients did not form a semi-circular shape.

## IRT Parameter Estimates

### *Item Locations*

The second research question focused on the item location estimates generated from the three IRT models, as it was hypothesized that locations may be very different across the two types of IRT models (cumulative, unfolding), if the survey was constructed using a method that assumed a dominance response process. To investigate this, IRT calibrations were performed on all 10 Empowerment data sets with the application of the PCM, GPCM, and GGUM models. Item parameter estimates, including the location estimates, as presented in this section.

The items as they appear on the NCTWC survey are presented in Table 4.7. The underlying trait, Empowerment, and the respective scale upon which both item and person estimates are located are different for the cumulative (PCM, GPCM) and GGUM models. PCM and GPCM estimates may be interpreted similarly, though cannot be directly interpreted relative to the GGUM parameter estimates, both item and person. The average item location estimate and average standard error of the location estimates, aggregated across the 10 samples are presented in Table 4.8 for the Empowerment scale. Item location ranks by IRT model are also tabled. All item location estimates generated from the 10 GGUM analyses on the Empowerment items were moderately ($\delta_i$ = 2.743) to highly extreme ($\delta_i$ =

5.048). Within unfolding analyses, the order of item locations and the content of items should be consistent. Item content, may, to some extent be associated with the extremity of item location in that item parameters are estimated relative to the location of people. Given that the location parameters from GGUM analyses are estimated and center the item relative to theta, the item locations can be interpreted as corresponding to the point on the latent trait where the average item response would lie. However, if a sample is relatively homogeneous in their attitude, say neutral, then moderately negative (or positive) items will appear extremely negative (or positive).

The relative extremity of the observed average item locations estimated from the GGUM could be an indication that the distribution of responses may be skewed in that many people agreed or strongly agreed to most of the items. Additionally, extreme estimates such as those presented in Table 4.8 could also result from scale drift that occurs when only a portion of the latent trait is measured. The signs of the item locations are arbitrary within unfolding analyses, therefore interpretation of an extreme location such as item 10 for example (average $\delta_i$ = 5.048) must include consideration of relative location of people, the items content and the content of the rest of the items on the scale. Item 10 reads: “Please indicate how large a role teachers at your school have in hiring new teachers” which arguably does not contain extreme content in either direction. Considering that the majority of respondents disagreed with this item, the extremity of the location of Item 10 is likely due to the homogeneity of attitudes among respondents with respect to this particular question. Table 4.9 displays the percentage of respondents who endorsed each category for each Empowerment item.

Table 4.7

*Empowerment Items (i = 13)*

Empowerment

Please rate your level of agreement with the following statements:

1. Teachers are centrally involved in decision making about educational issues
2. Teachers are trusted to make sound professional decisions about instruction
3. The faculty has an effective process for making group decisions and solving problems
4. In this school we take steps to solve problems
5. Opportunities for advancement within the teaching procession (other than administration) are available to me

Please indicate how large a role teachers at your school have in each of the following areas:

6. Selecting instructional materials and resources
7. Devising teaching techniques
8. Setting grading and student assessment practices
9. Determining the content of in-service professional development programs
10. Hiring new teachers
11. Establishing and implementing policies about student discipline
12. Deciding how the school budget will be spent
13. School improvement planning

Within the PCM and GPCM analyses, the majority of the item location estimates produced from the PCM and GPCM analyses were moderately negative, except for items 9 through 12, which were consistently estimated to be positive. The scale for item parameter estimates (and person parameter estimates) within IRT analyses typically ranges between -3 and +3. For example, application of the PCM for item 7 resulted in an average estimated item location value of -.936 which indicates that it was generally easy for respondents to endorse item 7. Alternatively, the average location estimate for item 10 resulting from the application of the GPCM was 2.39; this value is interpreted to indicate that a relatively good or positive attitude toward Empowerment is required for respondents to endorse this item. Based on these results, item 10 can be interpreted as difficult to agree with, or endorse. Put another way, it would take a very high degree of or very positive attitude about teacher

Empowerment to endorse item 10. The PCM and GPCM results are similar to each other at the item level in terms of both the location and standard error of the estimates. Additionally, the order of the location of items on the latent trait was almost identical between the PCM and GPCM analyses.

The large standard errors associated with the GGUM estimates for the majority of the Empowerment items would generally be expected for such extreme item location estimates. The large standard errors may be an indication that these items are not located in the same general region on the unidimensional latent trait as the majority of the respondents (i.e., thetas). Further, the general average location of the Empowerment items according to the GGUM analyses indicate that the items are located in one region of the latent trait scale, and not dispersed across the spectrum of the latent trait.

The average correlation between the Kendall's Tau-*b* parameter estimates across the 10 samples between the PCM and GPCM for estimated item locations was .938. The average correlation between the PCM and GGUM was .231 and the correlation between GPCM and GGUM item estimates was .180.  Across the 10 Empowerment samples, all correlations between the PCM and GPCM location estimates were statistically significant ($p < .05$). None of the 10 correlations between the PCM and GGUM location estimates were statistically significant, and none correlations between the GPCM and GGUM estimates were statistically significant. These correlations and associated *p* values, along with the nearly identical rank ordering of item locations are further evidence that the PCM and GPCM function similarly, as expected.

Table 4.8

*Average Item Locations, Standard Errors, and Rank Order of Item Locations across 10 Samples: Empowerment*

| | Average Item Location (Standard Error) | | | Rank Order of Average Item Locations | | |
|---|---|---|---|---|---|---|
| Item | PCM | GPCM | GGUM | PCM | GPCM | GGUM |
| 7 | -0.936 | -1.034 | -4.613 | 13 | 13 | 3 |
| | (0.039) | (0.048) | (3.735) | | | |
| 2 | -0.633 | -0.574 | -2.795 | 12 | 10 | 11 |
| | (0.038) | (0.031) | (0.727) | | | |
| 6 | -0.627 | -0.671 | -4.301 | 11 | 12 | 6 |
| | (0.037) | (0.042) | (3.181) | | | |
| 4 | -0.499 | -0.441 | -2.743 | 10 | 9 | 12 |
| | (0.041) | (0.030) | (0.542) | | | |
| 8 | -0.436 | -0.589 | -4.994 | 9 | 11 | 2 |
| | (0.036) | (0.057) | (5.030) | | | |
| 3 | -0.278 | -0.255 | -2.847 | 8 | 7 | 10 |
| | (0.041) | (0.028) | (0.743) | | | |
| 5 | -0.274 | -0.289 | -3.769 | 7 | 8 | 9 |
| | (0.035) | (0.046) | (1.861) | | | |
| 1 | -0.110 | -0.120 | -2.738 | 6 | 6 | 13 |
| | (0.038) | (0.028) | (0.463) | | | |
| 13 | -0.037 | -0.034 | -3.836 | 5 | 5 | 8 |
| | (0.036) | (0.035) | (2.429) | | | |
| 11 | 0.670 | 0.691 | -4.328 | 4 | 4 | 5 |
| | (0.035) | (0.036) | (3.921) | | | |
| 9 | 0.695 | 0.724 | -4.097 | 3 | 3 | 7 |
| | (0.035) | (0.037) | (3.374) | | | |
| 12 | 1.543 | 1.708 | -4.473 | 2 | 2 | 4 |
| | (0.037) | (0.047) | (6.601) | | | |
| 10 | 1.823 | 2.390 | -5.048 | 1 | 1 | 1 |
| | (0.038) | (0.072) | (10.601) | | | |

*Note:* The parameter estimates for the PCM and GPCM results are not directly comparable to the GGUM estimates as theta and the resulting scale are different.

The item location estimates, and the correlation between IRT models, reveal that the cumulative IRT models rank ordered the items differently than the unfolding model. However, none of the models estimated items to be evenly distributed across the latent trait.

Table 4.9

*Percentage of Category Endorsement by Empowerment Item: Full Sample (n = 65,008)*

| | Item | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Category | Strongly Disagree | 8.1 | 5.4 | 6.8 | 5.1 | 7.1 | 2.9 | 2.5 | 6.1 | 16.9 | 50 | 18.7 | 37.9 | 7 |
| | Disagre | 22.5 | 12.9 | 19.3 | 14.4 | 16.1 | 15 | 9.5 | 15.1 | 28.6 | 26.5 | 28.9 | 32.1 | 20.6 |
| | Neither Agree Nor Disagree | 17.8 | 10.3 | 17.9 | 17.1 | 23.7 | 31.6 | 24.3 | 26.7 | 32.3 | 15.3 | 28.5 | 20.3 | 31.7 |
| | Agree | 42.8 | 54.5 | 45.5 | 49.6 | 42.7 | 36.1 | 41.5 | 35.5 | 18.8 | 6.5 | 20.1 | 8.2 | 33.4 |
| | Strongly Agree | 8 | 16.2 | 9.9 | 12.5 | 10.1 | 13.5 | 21 | 15.4 | 2.6 | 0.6 | 3.1 | 0.9 | 6.6 |

The discriminating characteristics of the Empowerment items are described here for the three IRT models. Presented in Table 4.10 are the average item discriminations (*a* parameters) and standard errors across the 10 samples. As shown in Table 4.10, the values of the discrimination parameters generally range from 0 (no discrimination among examinees) to about 3.0 (item discriminates well among examinees) and these parameters are comparable across the three IRT models. Because the PCM is a Rasch model, the discrimination parameter is a fixed item parameter.

Table 4.10

*Average Item Discrimination and Standard Errors across 10 Empowerment Samples*

| | Average Item Discrimination | | | Average Standard Error | | |
|---|---|---|---|---|---|---|
| Item | PCM | GPCM | GGUM | PCM | GPCM | GGUM |
| 1 | 0.626 | 1.373 | 2.349 | 0.005 | 0.051 | 0.102 |
| 2 | 0.626 | 1.010 | 1.710 | 0.005 | 0.032 | 0.072 |
| 3 | 0.626 | 1.433 | 2.451 | 0.005 | 0.055 | 0.107 |
| 4 | 0.626 | 1.228 | 2.088 | 0.005 | 0.045 | 0.089 |
| 5 | 0.626 | 0.383 | 0.647 | 0.005 | 0.015 | 0.033 |
| 6 | 0.626 | 0.531 | 0.891 | 0.005 | 0.020 | 0.051 |
| 7 | 0.626 | 0.488 | 0.824 | 0.005 | 0.021 | 0.055 |
| 8 | 0.626 | 0.309 | 0.513 | 0.005 | 0.014 | 0.039 |
| 9 | 0.626 | 0.591 | 1.003 | 0.005 | 0.023 | 0.045 |
| 10 | 0.626 | 0.371 | 0.636 | 0.005 | 0.021 | 0.038 |
| 11 | 0.626 | 0.602 | 1.020 | 0.005 | 0.023 | 0.046 |
| 12 | 0.626 | 0.513 | 0.875 | 0.005 | 0.025 | 0.044 |
| 13 | 0.626 | 0.673 | 1.138 | 0.005 | 0.024 | 0.052 |

The PCM analyses were run using the PARSCALE software, where item discrimination parameters were necessarily fixed. The constraints imposed on these parameters included a mean of 1.0 and a standard deviation of .0001. A real value prior mean of 1 is generally an accepted value when constraining discrimination parameters. Larger values of the standard deviation, such as .01, are not tight and should be decreased to contribute to constraining the slopes to 1.0. A tighter and smaller value for the standard deviation like .0001 was necessary in the current investigation. In the applied literature where command files are available, values as small as .0000001 are used as the standard deviation of the discrimination parameters for Rasch type IRT models (Kang & Chen, 2008). The value of .0000001 was used on a single Empowerment sample, and had no impact on item level fit statistics, number of E-M iterations, and the discrimination parameter increased by only .002. Therefore, the value of .0001 was deemed appropriate and sufficient in this investigation.

Across the PCM analyses on the 10 Empowerment samples, each took about 50 iterations of the E-M cycle to converge. Generally, many iterations required for convergence is indicative of some problem or a potentially ill-fitting model. The relatively large number of E-M iterations coupled with the values of .626 for the *a* parameter, indicate that, given the data and the model, forcing the *a* parameters to a distribution of (1, .0001) was difficult and that the PCM does not fit the data well. Although the *a* parameters did not match completely, the GPCM and GGUM analyses ordered the 13 items identically in terms of most to least discriminating. For example, both models estimated that item 3 was most discriminating and that item 8 was least discriminating. The average Kendall’s Tau-*b* correlation between the GPCM and GGUM models, across the 10 Empowerment samples was .995. These correlations, across the 10 samples were all statistically significant ($p < .05$).

Because the majority of the sample either agreed or strongly agreed with item 10, this item also was associated with a low discrimination parameter. This item did not discriminate well among respondents, as most people agreed regardless of their standing on the latent trait.

Items 1 and 3, for example had relatively high discrimination parameter estimates, meaning that these items differentiate well between those respondents with low and high levels of the latent trait, attitude towards teacher Empowerment. Another contributing factor to the high discrimination estimates was the use of all 5 response categories by respondents.

The final item parameters estimated using the IRT models were the category probability thresholds. Because there were five categories or response options on the NCTWCS, four threshold parameters were estimated for each item. Average category threshold parameters across the 10 samples are presented in Table 4.11 for the PCM and GPCM models. The average values in Table 4.11 are denoted $d_k$ (*k* representing item category) and often referred to as threshold parameters. Within the PCM and GPCM specifications, $b_{j1}$ and $d_1$ are always equal to 0. According to the PCM and GPCM, in order to identify the point of intersection on the latent trait scale, between the probability of endorsing category 1 (Strongly Disagree) and for endorsing category 2 (Disagree), calculation of the item step parameters, $b_{jk}$, is necessary. Item step parameters are simply the difference between item location, $b_j$ and the threshold parameter, $d_k$.

Within the GGUM, somewhat analogous to the PCM and GPCM, are item category threshold parameters, denoted $\tau_{jk}$. Unlike the interpretation of the item step parameters within the PCM and GPCM models, the category threshold parameters within the GGUM denote the intersection of the subjective response category (SRC) functions relative to the item location. The threshold parameters within GGUM are not interpreted as the point of intersection of the observed response categories, although they are an indication of the variation across response options by the respondents. Table 4.12 presents the average PCM and GPCM item step parameters and the average GGUM item category threshold parameters.

Table 4.11

*Average Category Threshold Parameters Across 10 Empowerment Samples*

| | PCM | | | | GPCM | | | |
|---|---|---|---|---|---|---|---|---|
| Item | $d_2$ | $d_3$ | $d_4$ | $d_5$ | $d_2$ | $d_3$ | $d_4$ | $d_5$ |
| 1 | 1.792 | 0.072 | 0.588 | -2.452 | 1.414 | 0.273 | 0.167 | -1.853 |
| 2 | 1.393 | -0.168 | 1.047 | -2.272 | 1.245 | 0.048 | 0.587 | -1.881 |
| 3 | 1.710 | 0.154 | 0.540 | -2.404 | 1.363 | 0.313 | 0.129 | -1.804 |
| 4 | 1.605 | 0.245 | 0.529 | -2.379 | 1.352 | 0.335 | 0.174 | -1.862 |
| 5 | 1.483 | 0.624 | 0.227 | -2.334 | 1.722 | 0.772 | 0.578 | -3.072 |
| 6 | 2.162 | 0.586 | -0.510 | -2.239 | 2.305 | 0.624 | -0.516 | -2.413 |
| 7 | 1.705 | 0.638 | -0.278 | -2.065 | 1.828 | 0.709 | -0.241 | -2.296 |
| 8 | 1.488 | 0.591 | -0.174 | -1.905 | 1.915 | 0.829 | -0.029 | -2.714 |
| 9 | 1.851 | 0.910 | -0.334 | -2.427 | 1.891 | 0.960 | -0.323 | -2.527 |
| 10 | 1.415 | 0.890 | -0.076 | -2.228 | 1.550 | 1.214 | 0.113 | -2.876 |
| 11 | 1.717 | 0.713 | -0.129 | -2.301 | 1.741 | 0.749 | -0.109 | -2.381 |
| 12 | 1.688 | 0.833 | -0.199 | -2.323 | 1.800 | 0.940 | -0.174 | -2.567 |
| 13 | 1.916 | 0.722 | -0.198 | -2.440 | 1.858 | 0.714 | -0.185 | -2.387 |

Notes: $d_2$ = threshold parameter for category 2 (Disagree); $d_3$ = threshold parameter for category 3 (Neither Agree Nor Disagree); $d_4$ = threshold parameter for category 4 (Agree); $d_5$ = threshold parameter for category 5 (Strongly Agree)

Table 4.12

*Average Category Step and Threshold Parameters across 10 Empowerment Samples*

| | PCM | | | | GPCM | | | | GGUM | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Item | $b_{j2}$ | $b_{j3}$ | $b_{j4}$ | $b_{j5}$ | $b_{j2}$ | $b_{j3}$ | $b_{j4}$ | $b_{j5}$ | $\tau_{i2}$ | $\tau_{i3}$ | $\tau_{i4}$ | $\tau_{i5}$ |
| 1 | -1.902 | -.182 | -.698 | 2.342 | -1.533 | -.393 | -.286 | 1.733 | -4.277 | -3.138 | -3.030 | -.999 |
| 2 | -2.026 | -.465 | -1.679 | 1.639 | -1.820 | -.623 | -1.161 | 1.306 | -4.638 | -3.400 | -3.978 | -1.471 |
| 3 | -1.989 | -.432 | -.818 | 2.125 | -1.617 | -.568 | -.384 | 1.549 | -4.475 | -3.424 | -3.228 | -1.293 |
| 4 | -2.104 | -.744 | -1.028 | 1.880 | -1.793 | -.776 | -.615 | 1.422 | -4.548 | -3.527 | -3.369 | -1.310 |
| 5 | -1.757 | -.898 | -.501 | 2.060 | -2.010 | -1.061 | -.867 | 2.783 | -5.790 | -4.845 | -4.650 | -.759 |
| 6 | -2.789 | -1.214 | -.117 | 1.611 | -2.976 | -1.295 | -.155 | 1.742 | -7.291 | -5.610 | -4.475 | -2.486 |
| 7 | -2.641 | -1.574 | -.658 | 1.130 | -2.862 | -1.743 | -.793 | 1.262 | -7.479 | -6.376 | -5.435 | -3.282 |
| 8 | -1.924 | -1.027 | -.262 | 1.469 | -2.504 | -1.418 | -.560 | 2.125 | -7.593 | -6.455 | -5.580 | -2.592 |
| 9 | -1.156 | -.215 | 1.029 | 3.123 | -1.167 | -.236 | 1.047 | 3.251 | -5.282 | -4.341 | -3.064 | -.742 |
| 10 | .408 | .933 | 1.899 | 4.051 | .840 | 1.176 | 2.277 | 5.266 | -4.219 | -3.903 | -2.736 | .400 |
| 11 | -1.047 | -.043 | .799 | 2.971 | -1.051 | -.058 | .800 | 3.072 | -5.384 | -4.403 | -3.521 | -1.173 |
| 12 | -.145 | .710 | 1.742 | 3.866 | -.092 | .768 | 1.882 | 4.275 | -4.570 | -3.733 | -2.620 | .002 |
| 13 | -1.953 | -.759 | .162 | 2.403 | -1.892 | -.748 | .152 | 2.353 | -5.749 | -4.581 | -3.705 | -1.396 |

The average (across the 10 samples) points on the latent trait scale, where the category 1 (Strongly Disagree) and category 2 (Disagree) probabilities intersect within the PCM and GPCM models are given in the column labeled $b_{j2}$ . Likewise, the point of intersection of the category probabilities on the latent trait between the adjacent categories 2 (Disagree) and 3 (Neither Agree Nor Disagree) is found in the $b_{j3}$ column. The same interpretation is made for the points on the latent trait where the probability of selecting category 3 (Neither Agree Nor Disagree) and category 4 (Agree) is the same ($b_{j4}$ column) and for the point where the probability of selecting category 4 (Agree) and category (Strongly Agree) is the same ($b_{j5}$). For example, across the averaged parameters estimated within the PCM analyses, the point on the latent trait (Empowerment) scale, where the probability of endorsing category 1 (Strongly Disagree) and category 2 (Disagree) intersect, is located on average, at -1.902.

The similarities of the averaged step parameters between the PCM and GPCM are evident in that there is a general progression across categories. The locations on the latent trait within $b_{j2}$, $b_{j3}$, and $b_{j4}$ are relatively near each other, whereas a gap on the latent exists between those values and the values of $b_{j5}$. The relatively high positive values from both PCM and GPCM analyses within $b_{j5}$ can be interpreted as a relatively high or positive attitude required to strongly agree or agree with the Empowerment items. Items 9, 10, 11, and 12 exemplify this point. The clustering of the first three step parameters and the separation of the fourth step parameter may be an indication that respondents are not using the categories equally in that a disproportionate number of respondents are endorsing the lower (i.e., Strongly Disagree and Disagree) response options for items 9, 10, 11, and 12. These results are consistent with the item location parameter estimates, in that within the

PCM and GPCM analyses, eight of the 13 items were located between -1 and 0 on the latent trait scale. For those five items generally located between 0 and 2, the step parameters were also located in that interval on the latent trait. For example, the average location for item 10 within the PCM analyses was 1.823. The associated step parameters ranged between .408 and 4.051, suggesting that this item requires a very positive attitude towards the Empowerment of teachers, and that even those respondents who have a moderately positive attitude ($b_{j2}$ value = .408) are still likely to strongly disagree or disagree.

The threshold parameters within the GGUM analyses are not directly interpreted at the observed response level. Examination of the category probability plots is useful for the interpretation of item parameter estimates for GGUM analyses. Derivation and examination of the probability plots were the focus of the third research question, as it was hypothesized that, for the items that contained relatively neutral content, the plots would display characteristics of the ideal point response process (i.e., single-peaked, non-monotonic). It was also hypothesized that the two types of IRT models would function similarly if the attitudes possessed by the sample were located on one side of the items (i.e., homogeneous sample not measured well by items). Figures 4.2 through 4.14 display the category probability functions for the 13 Empowerment items from application of the PCM on the first simple random sample. Figures 4.15 through 4.27 display the category probability functions for the 13 Empowerment items from application of the GPCM on the first simple random sample, and Figures 4.28 through 4.40 display the category probability plots for the 13 Empowerment items resulting from the GGUM analyses on the first sample.

Figure 4.2

*Category Probability Plot for Item 1 with PCM: Sample 1, Empowerment Scale*

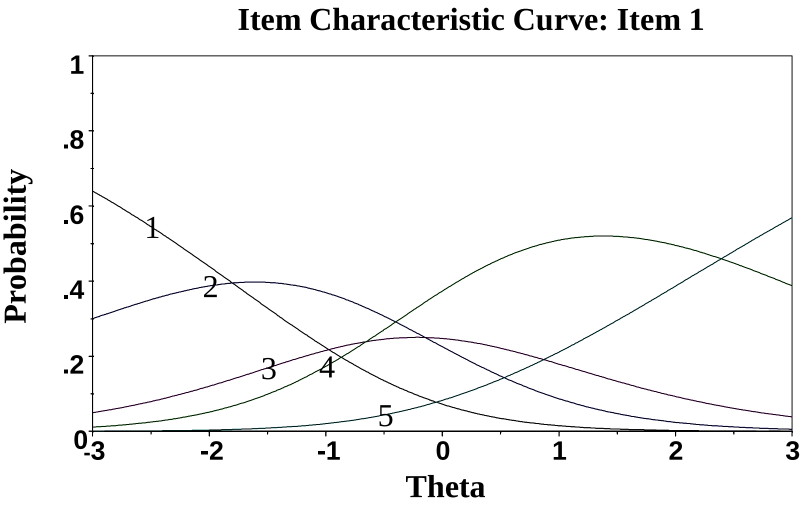


Figure 4.3

*Category Probability Plot for Item 2 with PCM: Sample 1, Empowerment Scale*

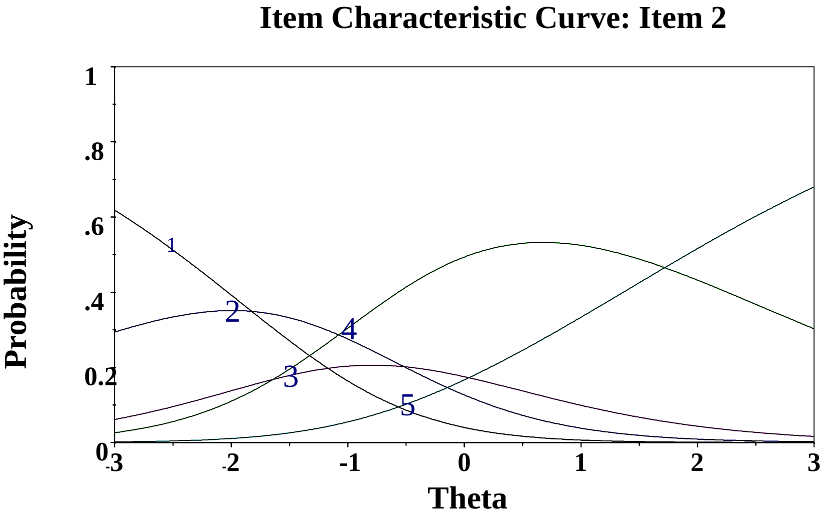


Figure 4.4

*Category Probability Plot for Item 3 with PCM: Sample 1, Empowerment Scale*

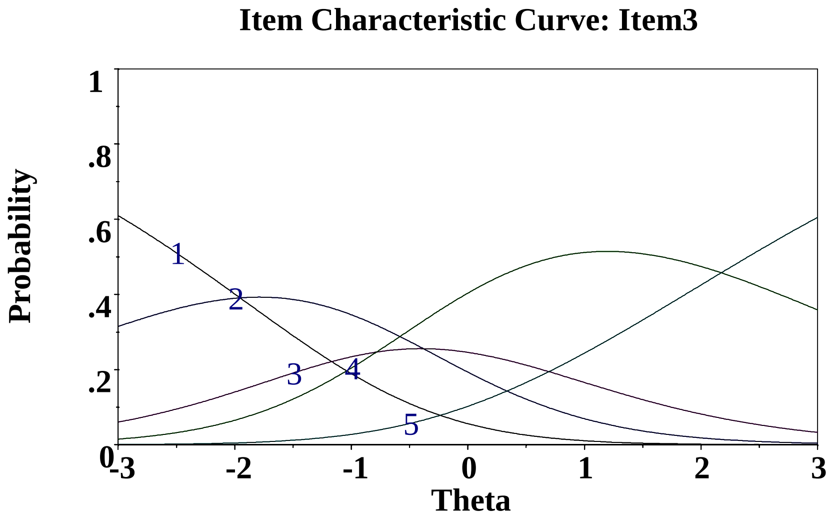


Figure 4.5

*Category Probability Plot for Item 4 with PCM: Sample 1, Empowerment Scale*

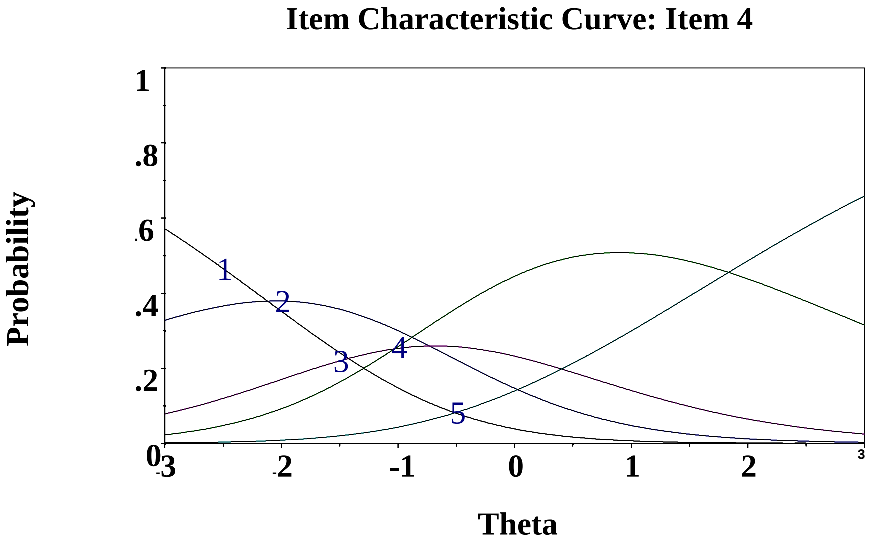


Figure 4.6

*Category Probability Plot for Item 5 with PCM: Sample 1, Empowerment Scale*

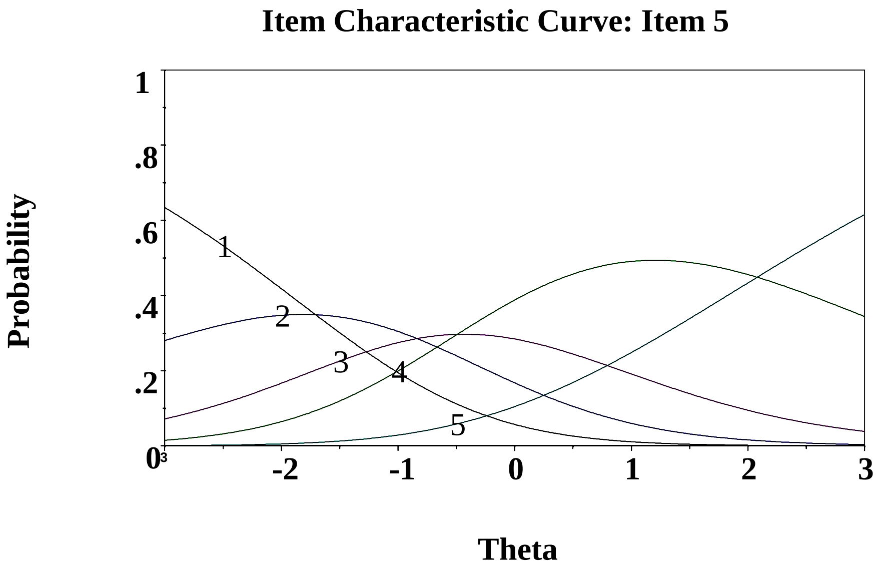


Figure 4.7

*Category Probability Plot for Item 6 with PCM: Sample 1, Empowerment Scale*

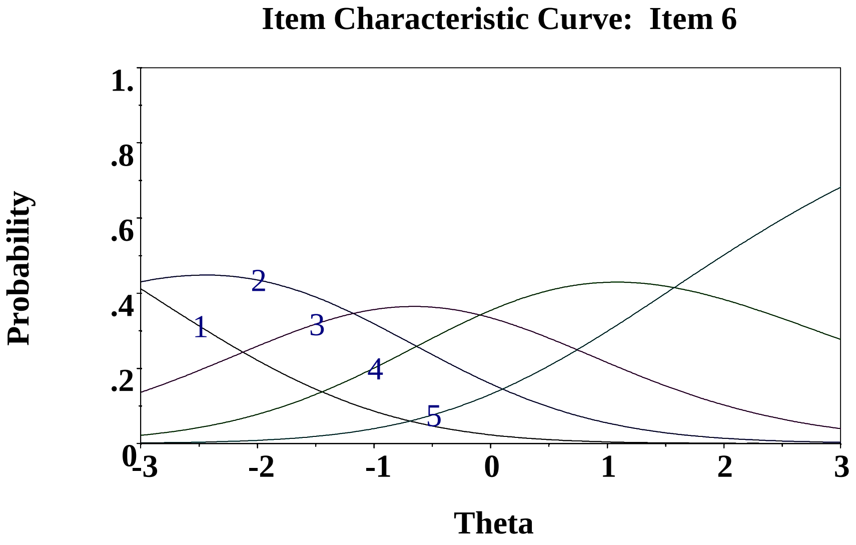


Figure 4.8

*Category Probability Plot for Item 7 with PCM: Sample 1, Empowerment Scale*

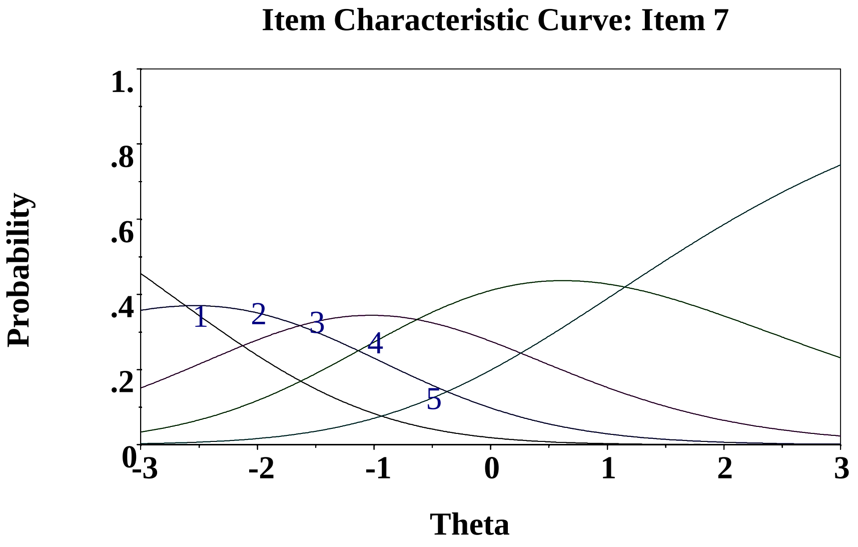


Figure 4.9

*Category Probability Plot for Item 8 with PCM: Sample 1, Empowerment Scale*

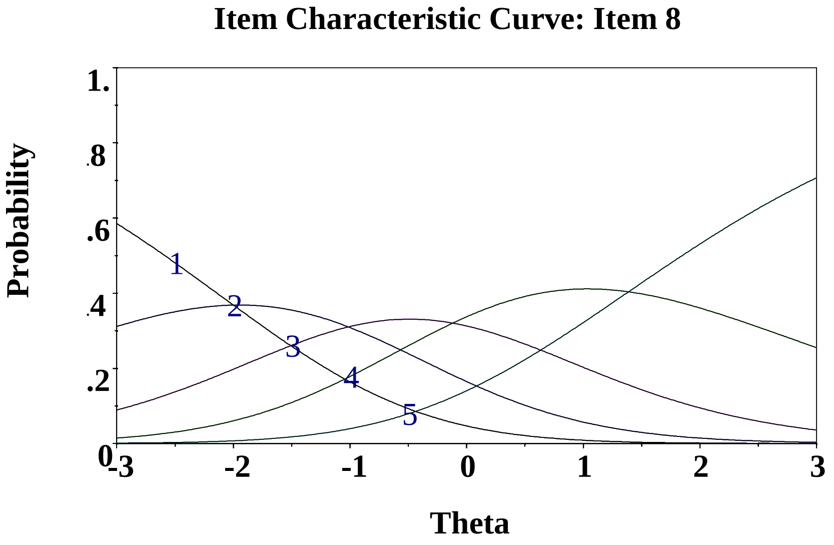


Figure 4.10

*Category Probability Plot for Item 9 with PCM: Sample 1, Empowerment Scale*

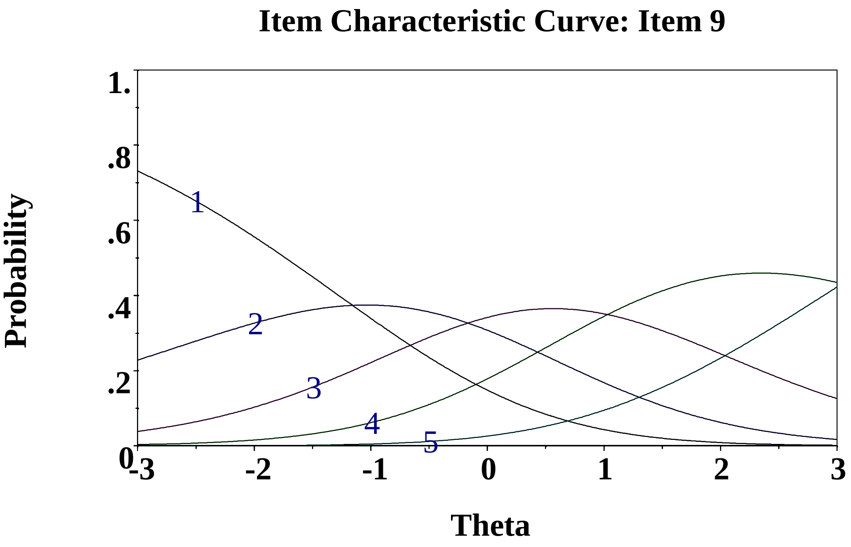


Figure 4.11

*Category Probability Plot for Item 10 with PCM: Sample 1, Empowerment Scale*

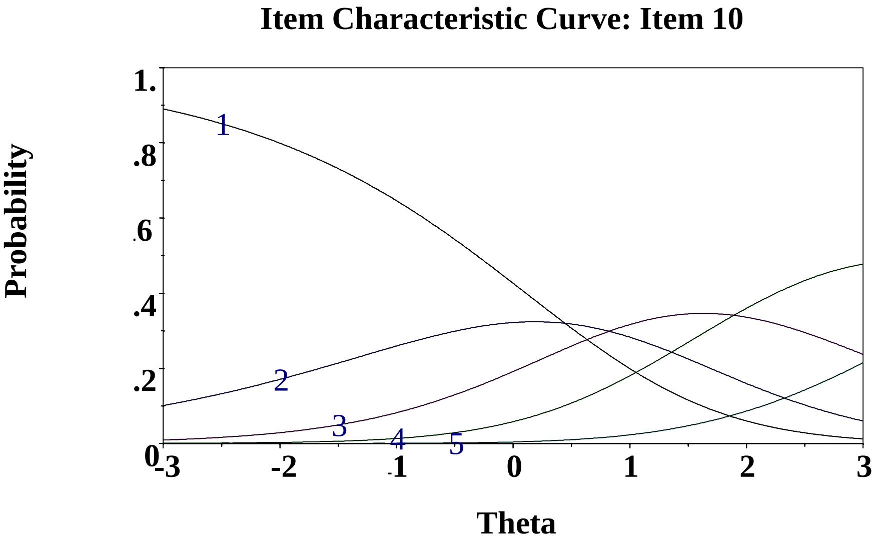


Figure 4.12

*Category Probability Plot for Item 11 with PCM: Sample 1, Empowerment Scale*

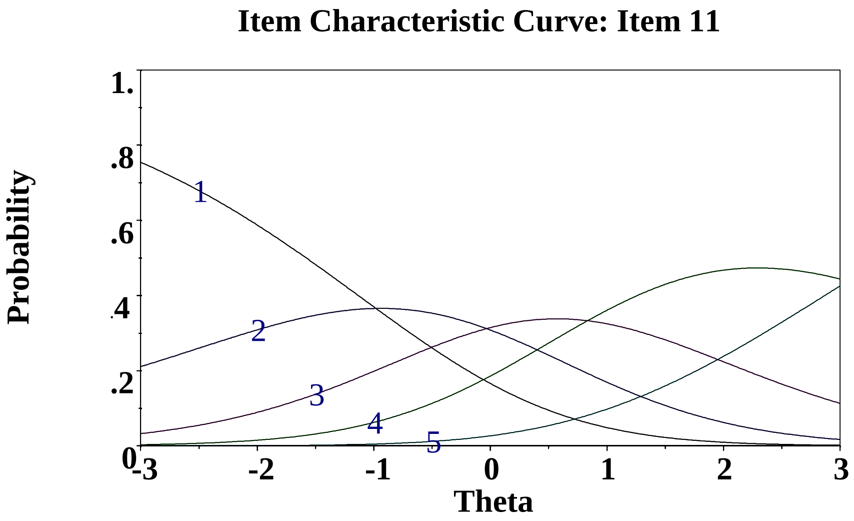


Figure 4.13

*Category Probability Plot for Item 12 with PCM: Sample 1, Empowerment Scale*

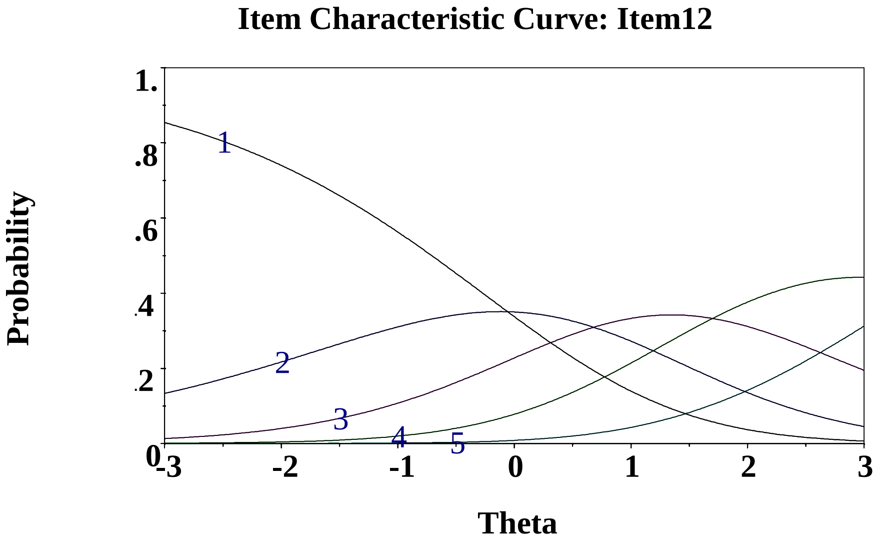


Figure 4.14

*Category Probability Plot for Item 13 with PCM: Sample 1, Empowerment Scale*

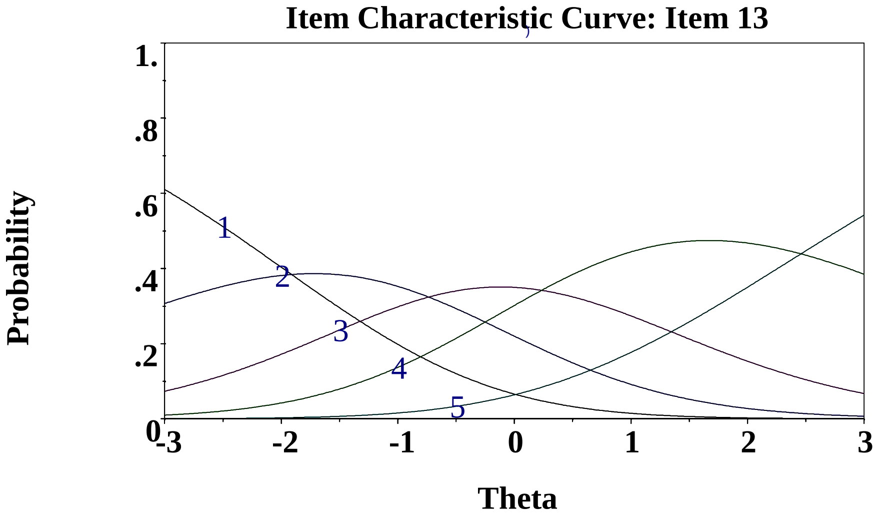


Figure 4.15

*Category Probability Plot for Item 1 with GPCM: Sample 1, Empowerment Scale*

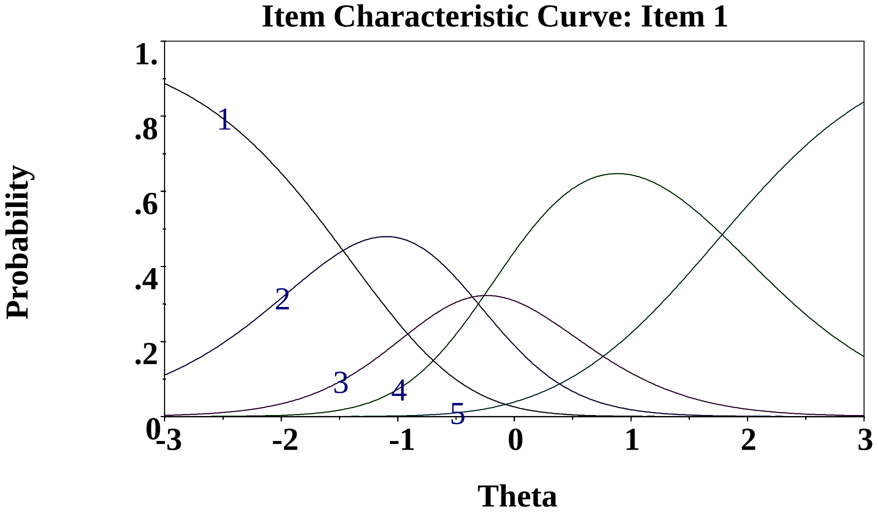


Figure 4.16

*Category Probability Plot for Item 2 with GPCM: Sample 1, Empowerment Scale*

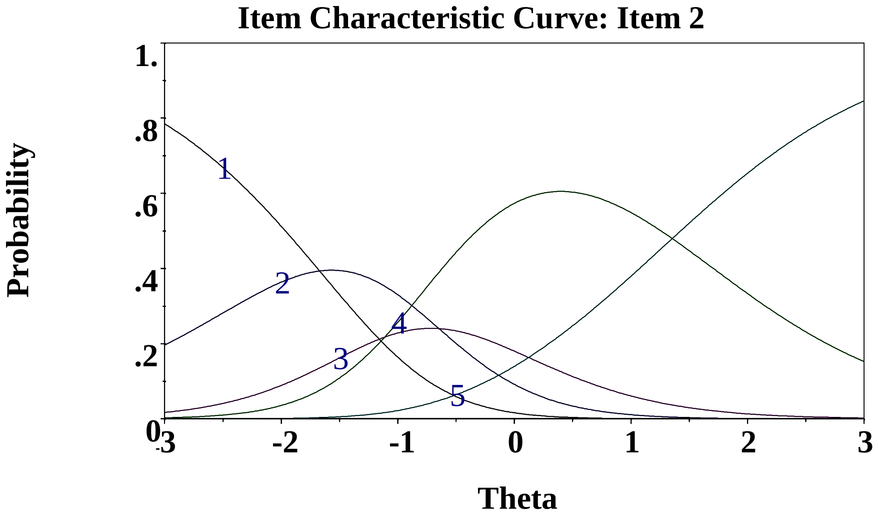


Figure 4.17

*Category Probability Plot for Item 3 with GPCM: Sample 1, Empowerment Scale*

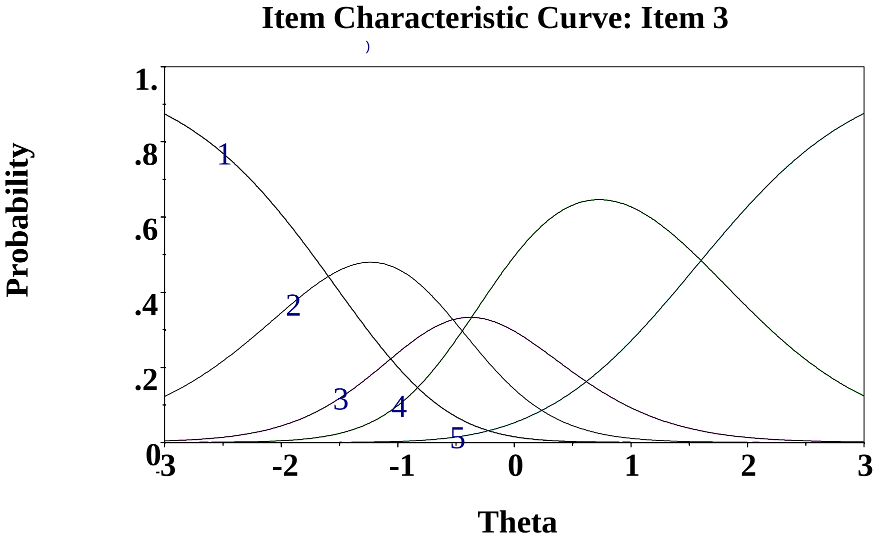


Figure 4.18

*Category Probability Plot for Item 4 with GPCM: Sample 1, Empowerment Scale*

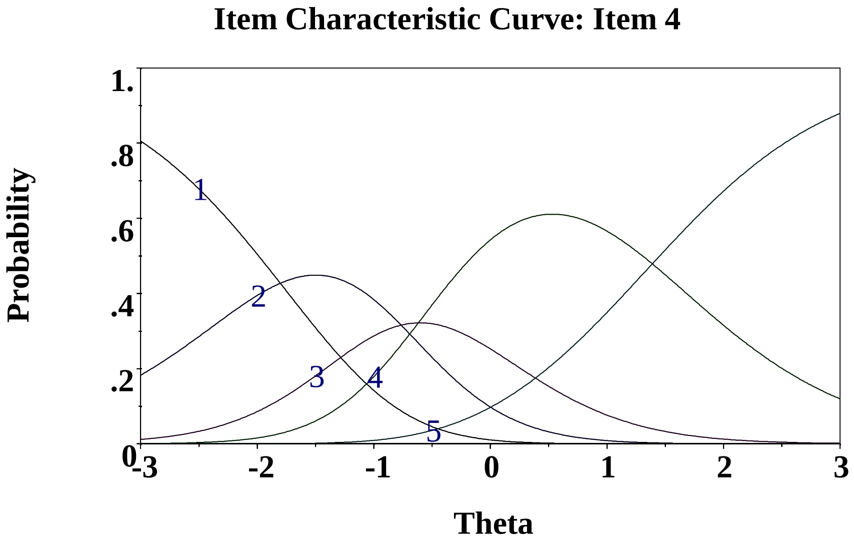


Figure 4.19

*Category Probability Plot for Item 5 with GPCM: Sample 1, Empowerment Scale*

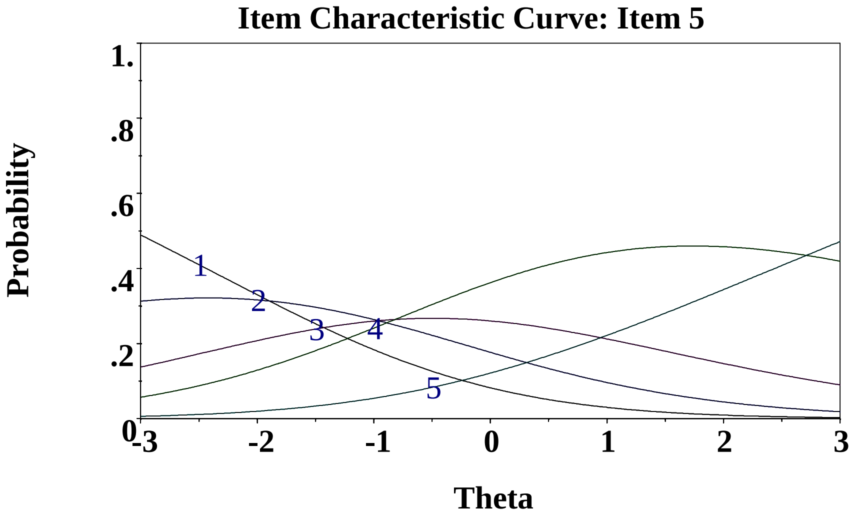


Figure 4.20

*Category Probability Plot for Item 6 with GPCM: Sample 1, Empowerment Scale*

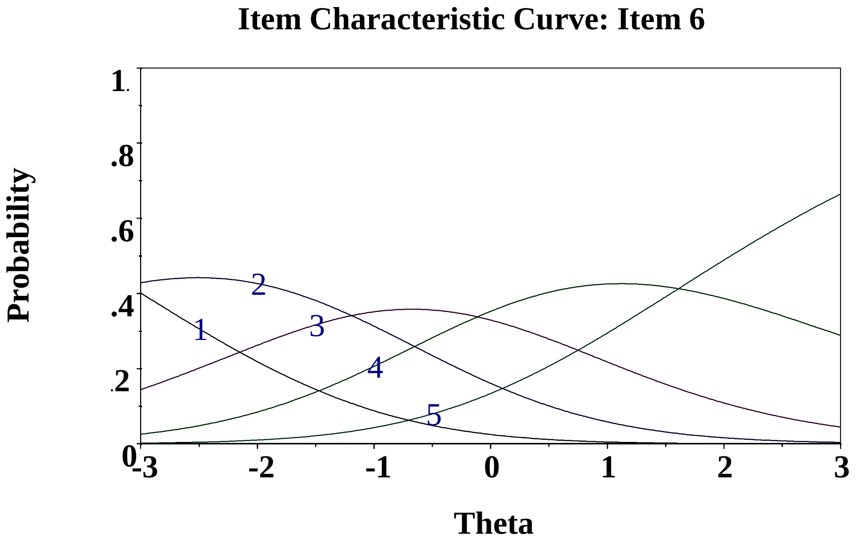


Figure 4.21

*Category Probability Plot for Item 7 with GPCM: Sample 1, Empowerment Scale*

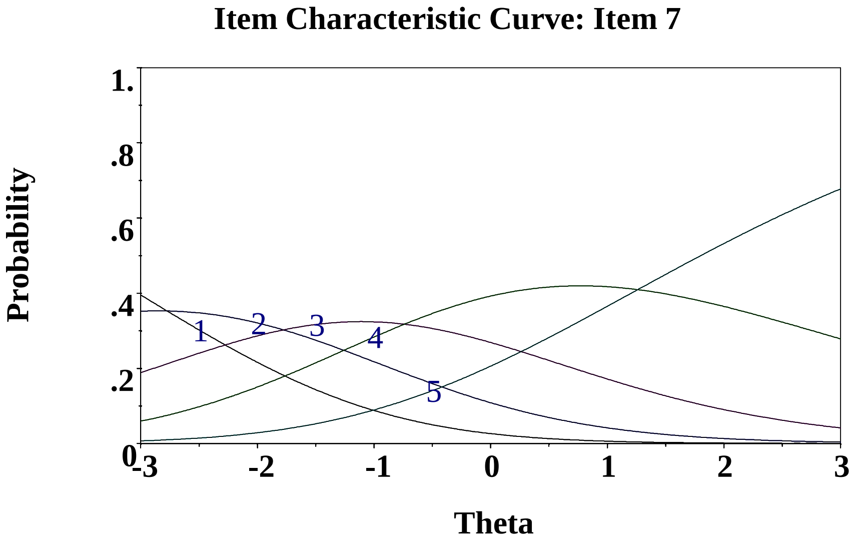


Figure 4.22

*Category Probability Plot for Item 8 with GPCM: Sample 1, Empowerment Scale*

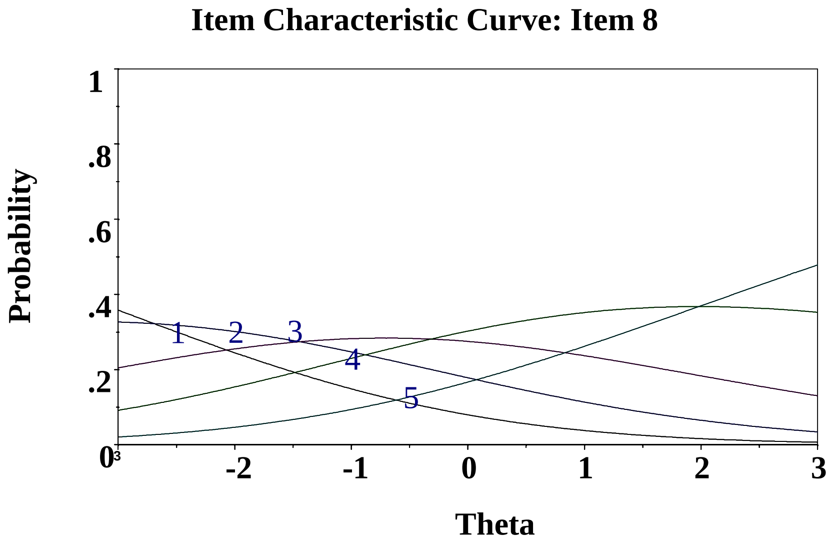


Figure 4.23

*Category Probability Plot for Item 9 with GPCM: Sample 1, Empowerment Scale*

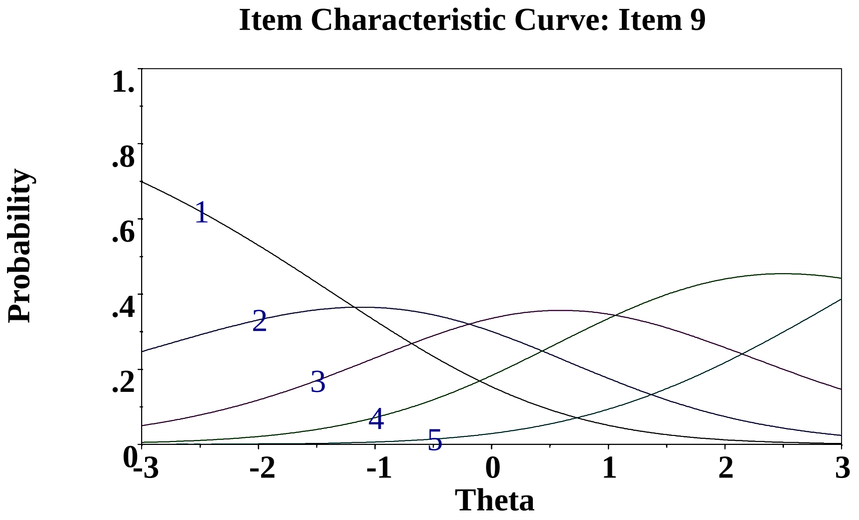


Figure 4.24

*Category Probability Plot for Item 10 with GPCM: Sample 1, Empowerment Scale*

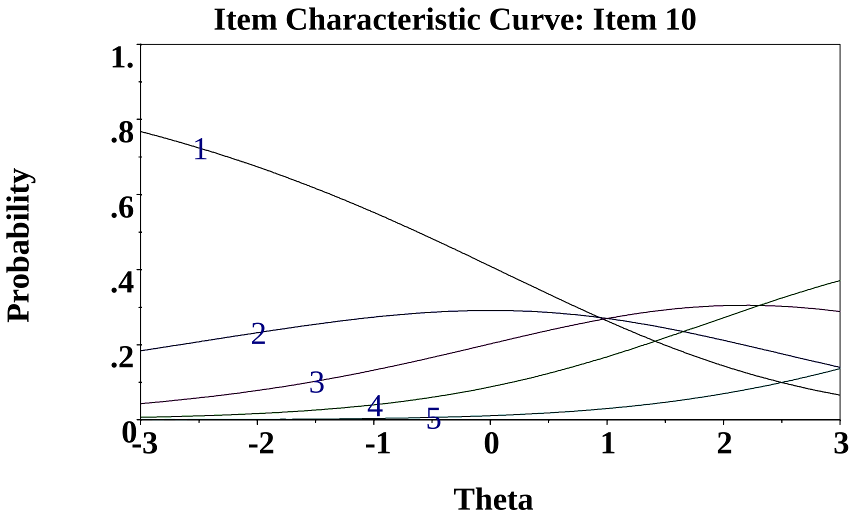


Figure 4.25

*Category Probability Plot for Item 11 with GPCM: Sample 1, Empowerment Scale*

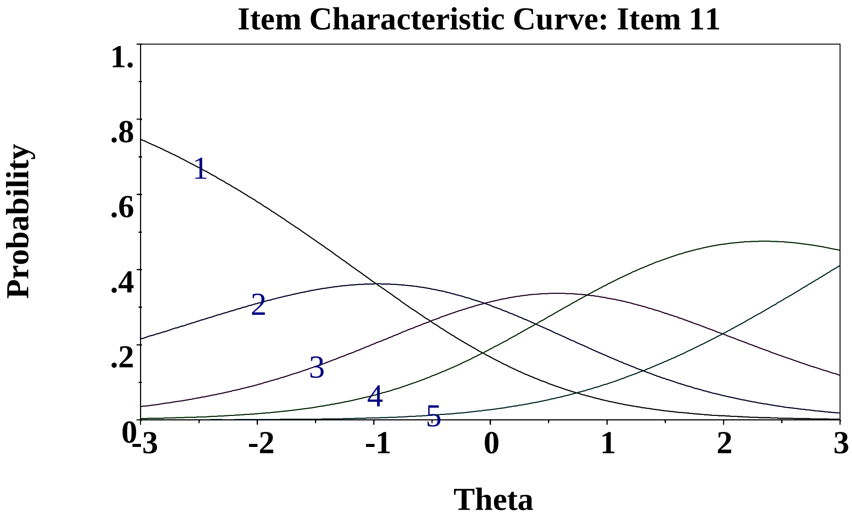


Figure 4.26

*Category Probability Plot for Item 12 with GPCM: Sample 1, Empowerment Scale*

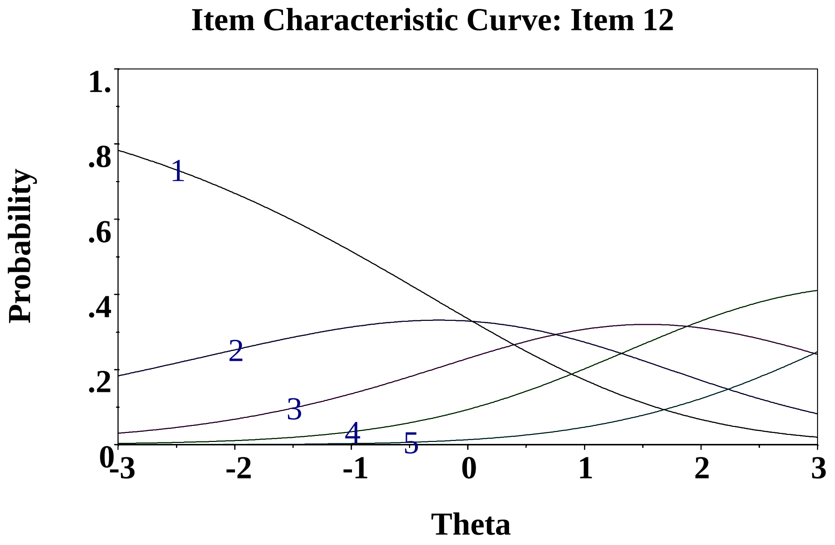


Figure 4.27

*Category Probability Plot for Item 13 with GPCM: Sample 1, Empowerment Scale*

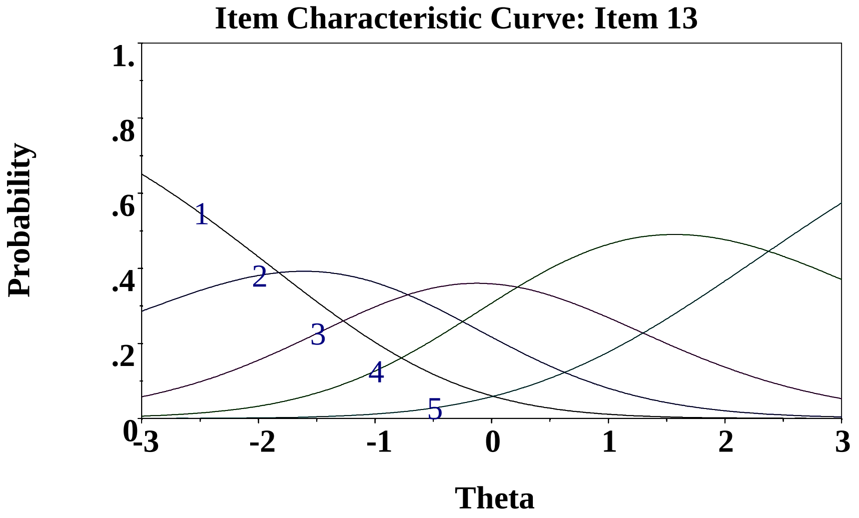


Figure 4.28

*Category Probability Plot for Item 1 with GGUM: Sample 1, Empowerment Scale*

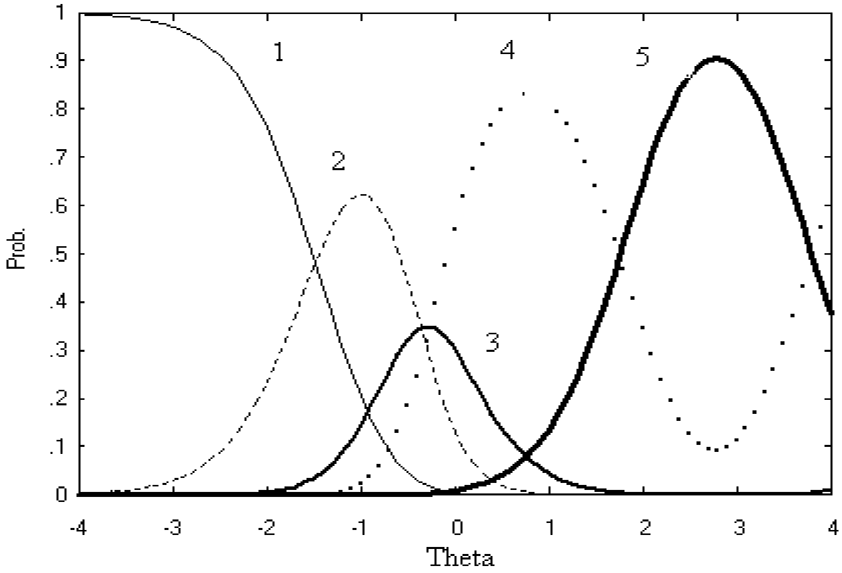


Figure 4.29

*Category Probability Plot for Item 2 with GGUM: Sample 1, Empowerment Scale*

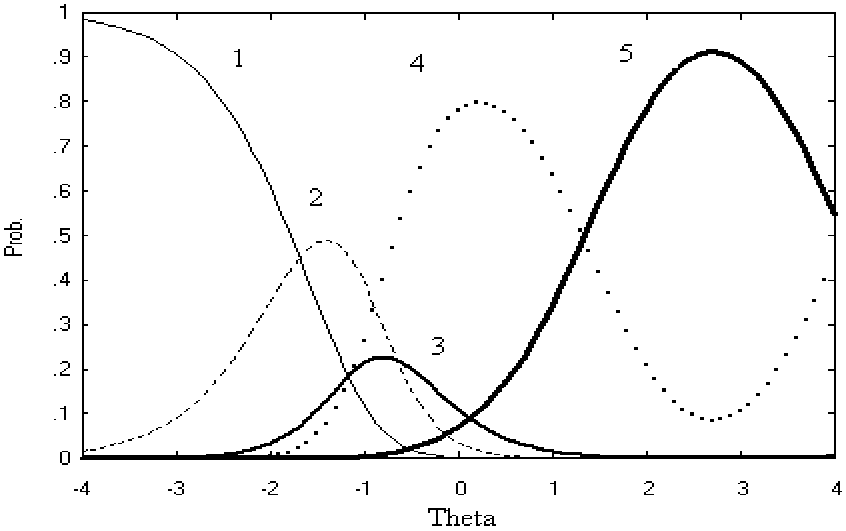


Figure 4.30

*Category Probability Plot for Item 3 with GGUM: Sample 1, Empowerment Scale*

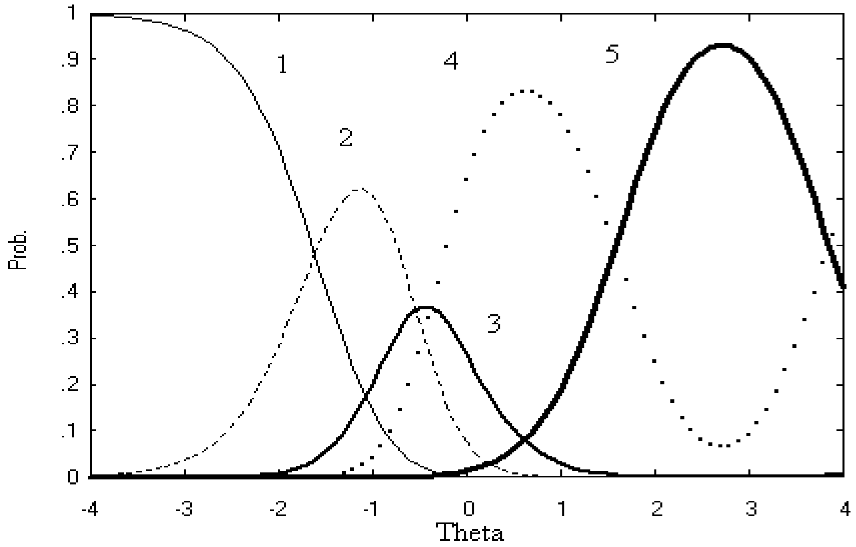


Figure 4.31

*Category Probability Plot for Item 4 with GGUM: Sample 1, Empowerment Scale*

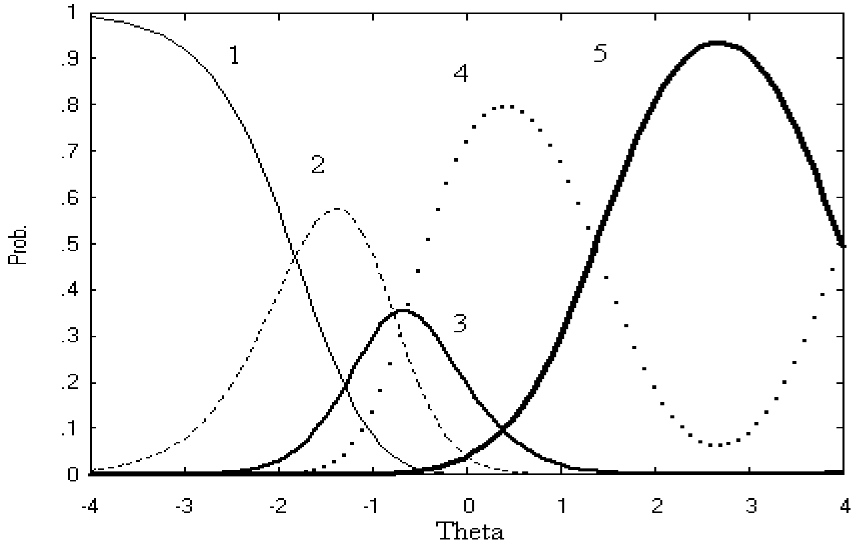


Figure 4.32

*Category Probability Plot for Item 5 with GGUM: Sample 1, Empowerment Scale*

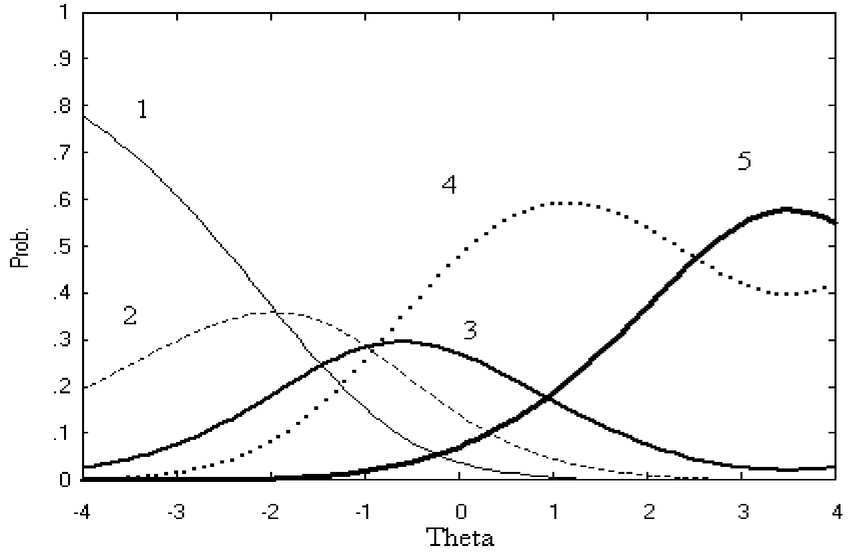


Figure 4.33

*Category Probability Plot for Item 6 with GGUM: Sample 1, Empowerment Scale*

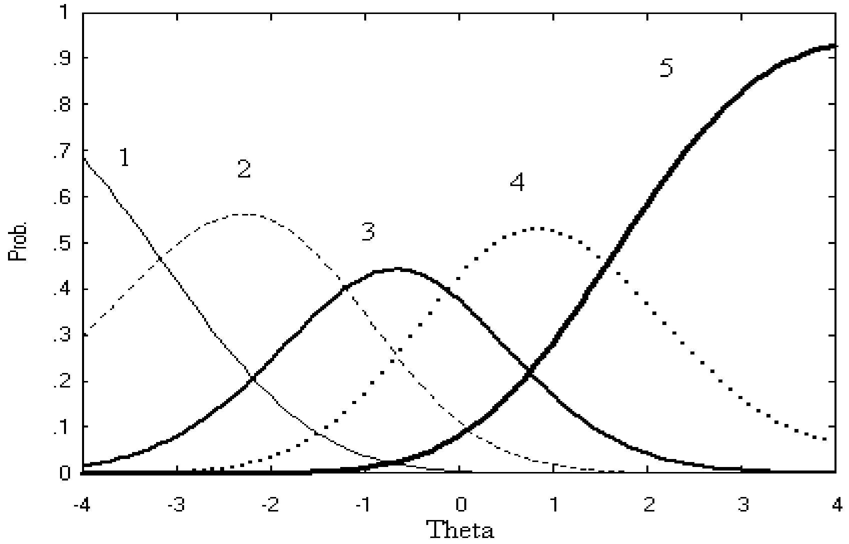


Figure 4.34

*Category Probability Plot for Item 7 with GGUM: Sample 1, Empowerment Scale*

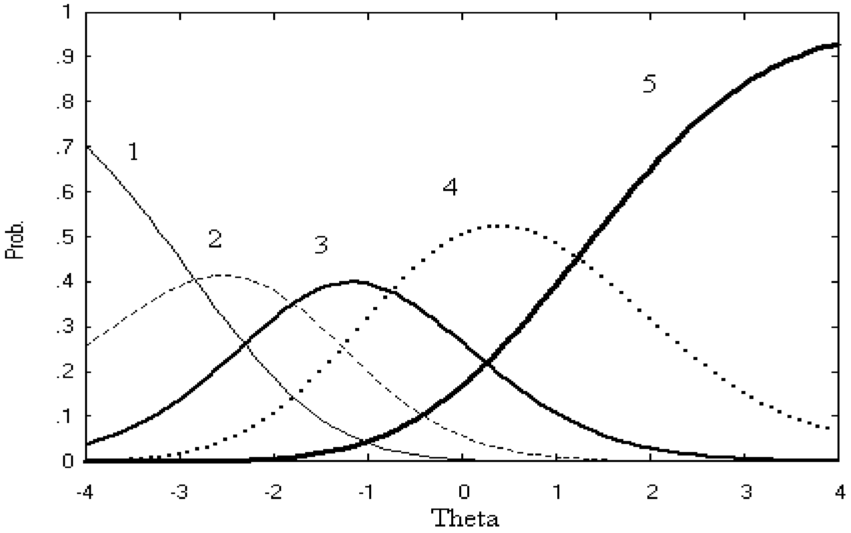


Figure 4.35

*Category Probability Plot for Item 8 with GGUM: Sample 1, Empowerment Scale*

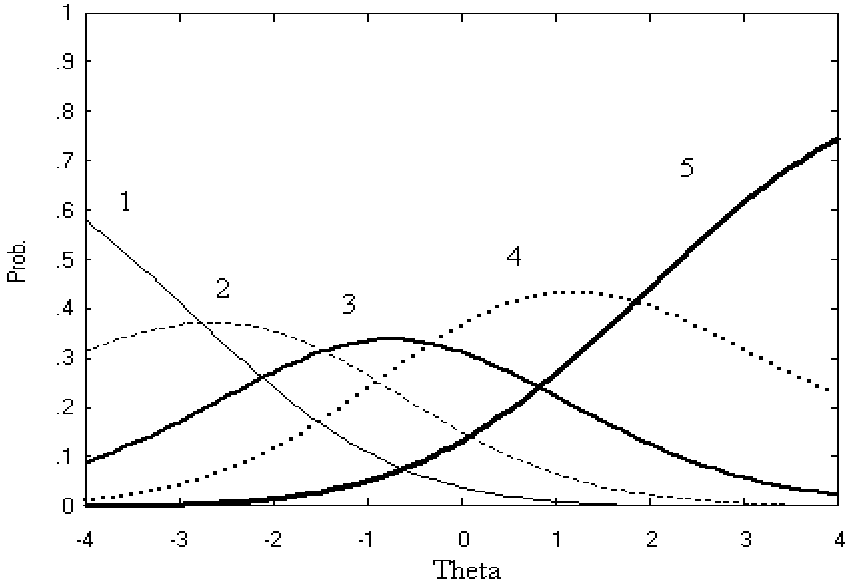


Figure 4.36

*Category Probability Plot for Item 9 with GGUM: Sample 1, Empowerment Scale*

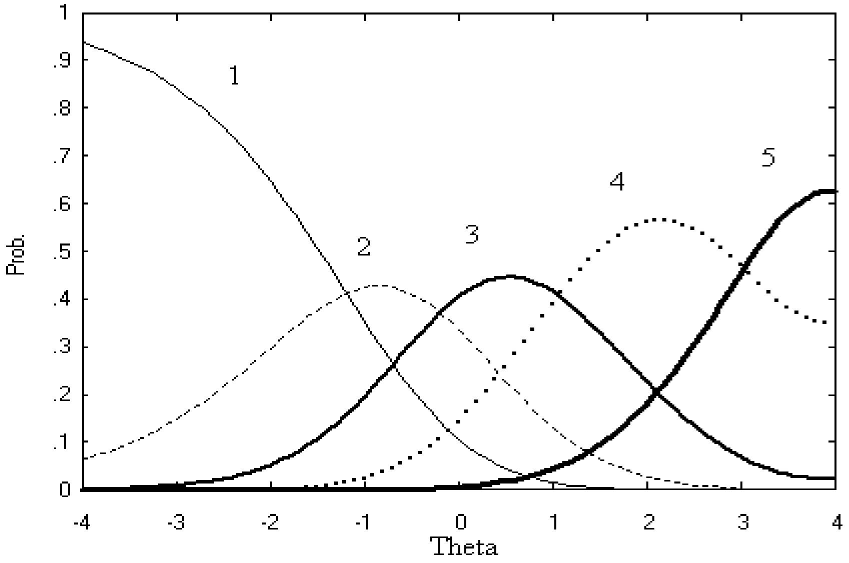


Figure 4.37

*Category Probability Plot for Item 10 with GGUM: Sample 1, Empowerment Scale*

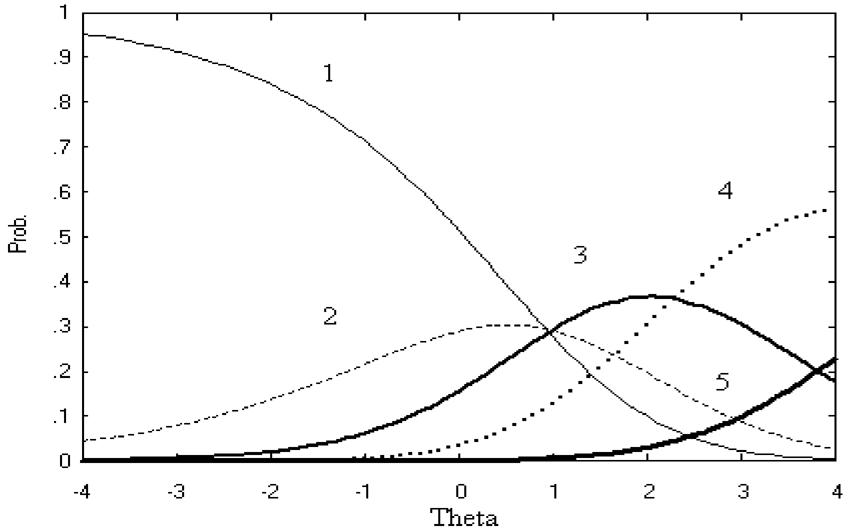


Figure 4.38

*Category Probability Plot for Item 11 with GGUM: Sample 1, Empowerment Scale*

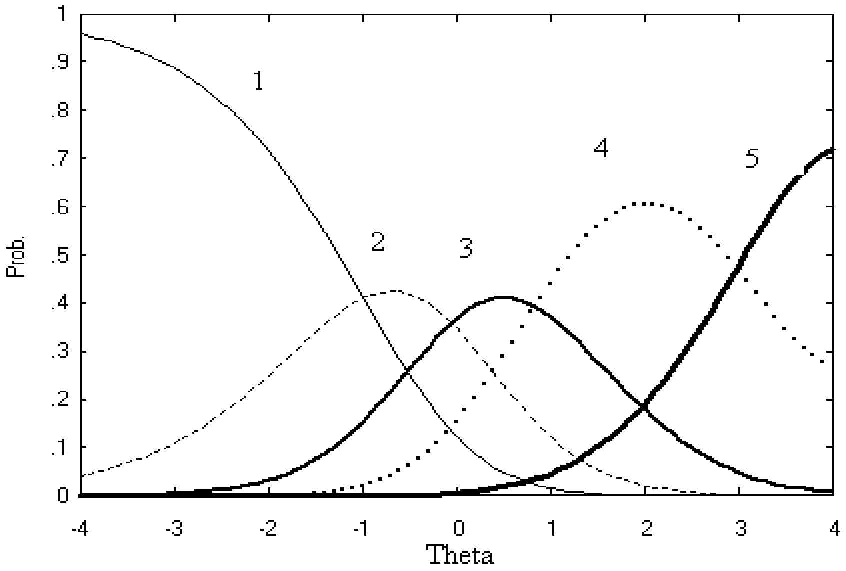


Figure 4.39

*Category Probability Plot for Item 12 with GGUM: Sample 1, Empowerment Scale*

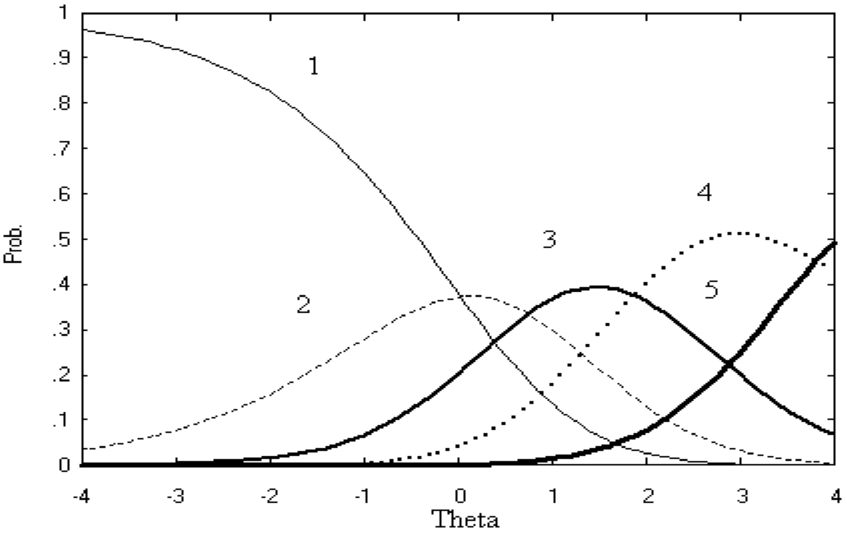
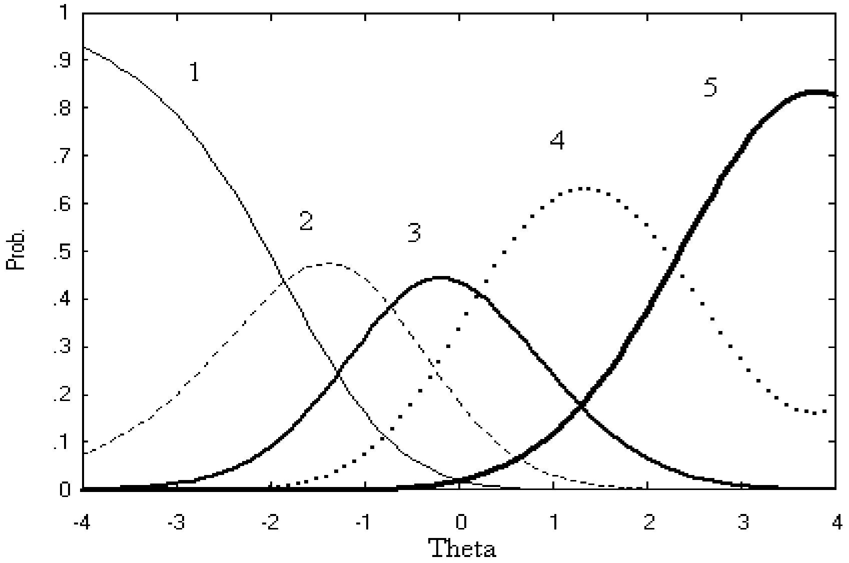


Figure 4.40

*Category Probability Plot for Item 13 with GGUM: Sample 1, Empowerment Scale*



To interpret the preceding figures, it is important to recall that within all category probability plots across the three IRT models, category 1 always represents Strongly Disagree and category 5 always represents Strongly Agree. These plots are graphical representations of the item parameters previously reported. For example, most respondents either agreed or strongly agreed with item 7 in sample 1 (see Figures 4.8, 4.21, 4.34 for PCM, GPCM, and GGUM, respectively). This item had a low discrimination parameter across all three IRT models, had a moderately negative average value for the location parameter (*b* = -.936, *b* = -1.034 for PCM and GPCM, respectively), meaning that this item was relatively easy to agree with, even for respondents with a moderately negative attitude towards teacher Empowerment. All of these components can be seen in the category probability plot of item 7, and for the rest of the items. Note that item discrimination is a constant or fixed parameter within the PCM, therefore only the location and step parameters affect the shape of the category probability plots within those analyses.

The category probability plots resulting from the GGUM analyses depict the same item characteristics. A characteristic of the GGUM model is that the observed category probability function of, say, Strongly Disagree is the summation of the probabilities associated with the two subject response category probabilities (Strongly Disagree from above and Strongly Disagree from below). The category response functions associated with the strongest level of agreement will peak around the point of the items estimated location. The response functions that represent Strongly Disagree and Disagree will peak furthest, in either direction, from the item's location. For example, Figure 4.37 depicts the category probability function resulting from the application of GGUM to item 10, in sample 1, an item with which very few respondents agreed or strongly agreed. According to the GGUM, this item displayed a low discrimination and was the most extreme item in terms of location ($\delta$ = 5.066) on the latent trait (Empowerment). The response function in Figure 4.37 associated with category 5 (Strongly Agree) is nearly non-existent because its peak would be located around the item location. Because the item location is so extreme and most people disagreed, the category response function associated with strongly disagree is monotonically decreasing. It can be determined for item 10, according to the GGUM, that respondents are disagreeing for one reason. A more discriminating (*a* = 2.363) and less extreme ($\delta$ = 2.271) item than item 10 in the first sample, according to GGUM is item 3. The response function for the Strongly Agree category peaks near the item's location. The highly discriminating nature of item 3 is evidenced by the more distinct and peaked category functions, compared to item 10.

Within the GGUM analyses, the Agree and Strongly Agree response categories for the first five items exhibit characteristics of an unfolding type item, or of an item to which the ideal point response process was used to answer. For example, an individual whose theta value is 3, is highly likely to strongly agree, and highly unlikely to agree with item 1. However, if an individual's value of theta is 2 or 4, then the probability of agreeing or strongly agreeing is approximately the same. This is characteristic of unfolding models; there are always two person estimates that yield the same probability of item endorsement. This pattern of probability is only evident for the categories of Agree and Strongly Agree and for the first five items. It is difficult to distinguish between the cumulative and unfolding models, based on the category plots for the rest of the items (6 through 13). The mean score for item 1 in sample 1 was 3.168, and the GGUM location for item1 in sample 1 was 2.779, therefore the category probability function that corresponds to Strongly Agree peaks around 3.

To address the third research question, the probability functions were generally very similar across the two types of IRT models, indicating little difference between cumulative and unfolding IRT models. Some items, did however, exhibit unfolding properties, especially for the Strongly Agree and Agree response options for the first four Empowerment items. These properties were also evidenced by the slight non-monotonicity of the ICCs generated from the GGUM analyses.

The second part of the fourth research question investigated in this study had to do with the fit of each model to the data. This was examined by calculating both absolute and relative fit statistics. Fit statistics are presented and discussed below. Both PARSCALE 4 and GGUM2004 calculate chi-square distributed (i.e., the likelihood ratio fit statistics ($G^2$) at the item and scale level. The PARSCALE 4 software collapses cells if frequencies are less than 5 (du Toit, 2003). Table 4.13 displays the item and scale (i.e., total) level fit statistics produced by the PCM for the 13 Empowerment items for each of the 10 samples. Table 4.14 contains the item and scale level fit statistics resulting from application of the GPCM, and Table 4.15 displays the item and scale level fit statistics estimated by GGUM. The null hypothesis for each of these tests is that observed and expected frequencies are the same across raw score groups (10 groups were specified for all analyses). The asterisks in Table 4.13, Table 4.14, and Table 4.15 denote those items that show good fit to the particular model (i.e., observed and expected cell frequencies do not differ statistically).

According to the PCM results, the only item that consistently displayed good model fit across all 10 samples was item 12 ("Please indicate how large a role teachers at your school have in deciding how the school budget will be spent"). Item 10, fit well in three samples. Fit of the PCM model for the whole Empowerment scale is given at the bottom of Table 4.13, where, across all 10 samples, the PCM does not fit these data well. The item and scale level chi-square distributed fit statistics for all 10 Empowerment Samples produced by the GPCM are given in Table 4.14. Within the GPCM analyses, only two items, 8 and 10 exhibited statistically good fit within more than half the samples. Item 8 reads: "Please indicate how large a role teachers at your school have in setting grading and student assessment practices" and item 10 reads "Please indicate how large a role teachers at your school have in hiring new teachers." Within the PCM analyses, item 12 consistently exhibited good statistical fit, where as in the GPCM analyses item 12 appeared to fit well in four samples. Similar to the PCM analyses, however, the GPCM did not appear to fit the Empowerment data, in any analysis at the item or scale level.

Table 4.13

*Item and Scale Level Chi-Square Fit Statistics for Each Empowerment Sample: PCM*

| | Sample 1 | | | Sample 2 | | | Sample 3 | | | Sample 4 | | | Sample 5 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Item | $\chi^2$ | *df* | *p* | $\chi^2$ | *df* | *p* | $\chi^2$ | *df* | *p* | $\chi^2$ | *df* | *p* | $\chi^2$ | *df* | *p* |
| 1 | 222.477 | 29 | .000 | 181.763 | 28 | .000 | 241.538 | 28 | .000 | 267.479 | 27 | .000 | 193.426 | 28 | .000 |
| 2 | 139.416 | 26 | .000 | 146.097 | 26 | .000 | 155.472 | 28 | .000 | 149.755 | 26 | .000 | 137.680 | 26 | .000 |
| 3 | 205.651 | 27 | .000 | 206.508 | 27 | .000 | 192.113 | 28 | .000 | 171.452 | 28 | .000 | 215.036 | 27 | .000 |
| 4 | 155.227 | 27 | .000 | 151.227 | 28 | .000 | 164.895 | 28 | .000 | 170.657 | 27 | .000 | 203.529 | 28 | .000 |
| 5 | 96.288 | 28 | .000 | 126.454 | 27 | .000 | 122.194 | 28 | .000 | 136.241 | 28 | .000 | 150.527 | 28 | .000 |
| 6 | 66.055 | 28 | .000 | 80.243 | 28 | .000 | 63.690 | 28 | .000 | 71.423 | 29 | .000 | 103.770 | 28 | .000 |
| 7 | 53.793 | 27 | .002 | 50.581 | 27 | .004 | 63.337 | 27 | .000 | 38.665 | 27 | .068* | 62.346 | 27 | .000 |
| 8 | 97.530 | 28 | .000 | 94.938 | 27 | .000 | 102.864 | 28 | .000 | 103.352 | 27 | .000 | 106.681 | 28 | .000 |
| 9 | 36.400 | 30 | .195* | 66.726 | 31 | .000 | 75.512 | 31 | .000 | 41.945 | 30 | .072* | 84.216 | 30 | .000 |
| 10 | 47.715 | 28 | .012 | 58.816 | 28 | .001 | 39.659 | 28 | .071* | 56.385 | 28 | .001 | 42.222 | 28 | .041 |
| 11 | 54.501 | 29 | .003 | 76.242 | 29 | .000 | 80.531 | 31 | .000 | 63.631 | 30 | .000 | 63.158 | 30 | .000 |
| 12 | 33.265 | 30 | .311* | 33.854 | 29 | .244* | 36.759 | 30 | .184* | 35.424 | 30 | .227* | 31.086 | 30 | .411* |
| 13 | 81.156 | 29 | .000 | 106.804 | 29 | .000 | 85.967 | 29 | .000 | 82.396 | 28 | .000 | 50.829 | 28 | .005 |
| Total | 1289.474 | 366 | .000 | 1380.254 | 364 | .000 | 1424.531 | 372 | .000 | 1388.804 | 365 | .000 | 1444.505 | 366 | .000 |

*Note:* * denotes observed and expected frequencies are not statistically different (α > .01)

Table 4.13 Con’t

*Item and Scale Level Chi-Square Fit Statistics for Each Empowerment Sample: PCM*

| | Sample 6 | | | Sample 7 | | | Sample 8 | | | Sample 9 | | | Sample 10 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Item | | *df* | *p* | | *df* | *p* | | *df* | *p* | | *df* | *p* | | *df* | *p* |
| 1 | 214.337 | 28 | .000 | 237.393 | 28 | .000 | 221.992 | 27 | .000 | 204.884 | 28 | .000 | 234.098 | 29 | .000 |
| 2 | 109.484 | 28 | .000 | 140.448 | 26 | .000 | 151.995 | 26 | .000 | 128.983 | 26 | .000 | 137.386 | 26 | .000 |
| 3 | 201.712 | 27 | .000 | 216.961 | 27 | .000 | 192.382 | 27 | .000 | 209.092 | 27 | .000 | 212.145 | 27 | .000 |
| 4 | 126.479 | 27 | .000 | 178.440 | 28 | .000 | 153.887 | 27 | .000 | 132.435 | 27 | .000 | 183.542 | 28 | .000 |
| 5 | 155.425 | 27 | .000 | 130.925 | 27 | .000 | 138.761 | 27 | .000 | 139.105 | 27 | .000 | 149.307 | 28 | .000 |
| 6 | 86.789 | 27 | .000 | 58.427 | 27 | .000 | 75.030 | 27 | .000 | 57.118 | 27 | .001 | 90.275 | 28 | .000 |
| 7 | 55.452 | 27 | .001 | 82.342 | 27 | .000 | 55.034 | 27 | .001 | 60.419 | 27 | .000 | 68.924 | 27 | .000 |
| 8 | 81.918 | 27 | .000 | 125.177 | 27 | .000 | 91.533 | 27 | .000 | 117.110 | 27 | .000 | 127.326 | 28 | .000 |
| 9 | 55.493 | 30 | .003 | 67.482 | 31 | .000 | 52.372 | 31 | .010 | 71.274 | 30 | .000 | 75.751 | 31 | .000 |
| 10 | 40.161 | 28 | .064* | 56.015 | 29 | .002 | 50.806 | 29 | .007 | 36.833 | 28 | .122* | 64.533 | 28 | .000 |
| 11 | 61.168 | 30 | .001 | 68.815 | 30 | .000 | 54.360 | 30 | .004 | 89.740 | 29 | .000 | 55.835 | 30 | .003 |
| 12 | 27.223 | 29 | .56* | 25.935 | 30 | .679* | 35.744 | 30 | .216* | 28.255 | 30 | .557* | 49.395 | 29 | .011 |
| 13 | 95.758 | 29 | .000 | 79.646 | 28 | .000 | 87.523 | 29 | .000 | 85.675 | 29 | .000 | 82.193 | 29 | .000 |
| Total | 1311.398 | 364 | .000 | 1468.006 | 365 | .000 | 1361.417 | 364 | .000 | 1360.922 | 362 | .000 | 1530.710 | 368 | .000 |

*Note:* * denotes observed and expected frequencies are not statistically different (α > .01)

The log-likelihood fit statistics ($G^2$) for each item on the Empowerment scale resulting from GGUM analyses are presented in Table 4.15.The *GGUM2004 Technical Reference Manual* (Roberts & Shim, 2008) cautions users of fit statistics:

> Users should be aware that these fit statistics and their associated degrees of freedom have been logically generalized (not mathematically deduced) from other cumulative IRT applications (which themselves may be suspect). Therefore, little is known about the distribution of these statistics, their Type I error rates, and their power rates under the GGUM (p. 34).

Table 4.14

*Item and Scale Level Chi-Square Fit Statistics for Each Empowerment Sample: GPCM*

| | Sample 1 | | | Sample 2 | | | Sample 3 | | | Sample 4 | | | Sample 5 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Item | $\chi^2$ | *df* | *p* | $\chi^2$ | *df* | *p* | $\chi^2$ | *df* | *p* | $\chi^2$ | *df* | *p* | $\chi^2$ | *df* | *p* |
| 1 | 59.747 | 23 | .000 | 51.903 | 24 | .001 | 50.115 | 22 | .001 | 58.615 | 20 | .000 | 44.587 | 22 | .003 |
| 2 | 79.990 | 24 | .000 | 100.276 | 24 | .000 | 92.049 | 23 | .000 | 66.907 | 21 | .000 | 108.578 | 23 | .000 |
| 3 | 69.357 | 22 | .000 | 63.561 | 23 | .000 | 73.339 | 22 | .000 | 68.522 | 22 | .000 | 48.427 | 21 | .001 |
| 4 | 61.722 | 23 | .000 | 60.492 | 23 | .000 | 84.358 | 22 | .000 | 68.821 | 22 | .000 | 50.140 | 22 | .001 |
| 5 | 83.258 | 32 | .000 | 135.977 | 31 | .000 | 105.417 | 32 | .000 | 142.810 | 31 | .000 | 140.929 | 31 | .000 |
| 6 | 73.863 | 28 | .000 | 65.388 | 29 | .000 | 97.055 | 29 | .000 | 72.392 | 28 | .000 | 114.250 | 28 | .000 |
| 7 | 64.884 | 27 | .000 | 57.880 | 28 | .001 | 73.728 | 28 | .000 | 50.485 | 27 | .004 | 67.883 | 27 | .000 |
| 8 | 36.321 | 33 | .316* | 30.354 | 34 | .647* | 55.740 | 32 | .006 | 32.562 | 31 | .39* | 52.389 | 31 | .010 |
| 9 | 45.770 | 31 | .042 | 41.519 | 31 | .098* | 45.280 | 30 | .036 | 52.521 | 30 | .007 | 65.329 | 30 | .000 |
| 10 | 43.765 | 33 | .099* | 55.350 | 33 | .009 | 45.629 | 35 | .108* | 74.550 | 34 | .000 | 36.896 | 33 | .293* |
| 11 | 49.243 | 30 | .015 | 65.172 | 31 | .000 | 82.921 | 30 | .000 | 85.662 | 31 | .000 | 84.327 | 32 | .000 |
| 12 | 66.028 | 32 | .000 | 75.153 | 32 | .000 | 63.228 | 33 | .001 | 42.612 | 32 | .099* | 50.810 | 31 | .014 |
| 13 | 50.684 | 28 | .005 | 77.495 | 28 | .000 | 90.662 | 30 | .000 | 44.017 | 28 | .028 | 49.043 | 28 | .008 |
| Total | 784.633 | 366 | .000 | 880.519 | 371 | .000 | 959.519 | 368 | .000 | 860.477 | 357 | .000 | 913.588 | 359 | .000 |

*Note*: * denotes observed and expected frequencies are not statistically different (α > .01)

Table 4.14 Con’t

*Item and Scale Level Chi-Square Fit Statistics for Each Empowerment Sample: GPCM*

| | Sample 6 | | | Sample 7 | | | Sample 8 | | | Sample 9 | | | Sample 10 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Item | $\chi^2$ | *df* | *p* | $\chi^2$ | *df* | *p* | $\chi^2$ | *df* | *p* | $\chi^2$ | *df* | *p* | $\chi^2$ | *df* | *p* |
| 1 | 56.737 | 22 | .000 | 42.049 | 22 | .006 | 52.885 | 22 | .000 | 51.380 | 21 | .000 | 48.290 | 22 | .001 |
| 2 | 81.814 | 24 | .000 | 66.261 | 21 | .000 | 65.430 | 21 | .000 | 82.936 | 23 | .000 | 58.871 | 24 | .000 |
| 3 | 50.205 | 22 | .001 | 45.184 | 20 | .001 | 49.819 | 22 | .001 | 64.011 | 21 | .000 | 45.934 | 21 | .001 |
| 4 | 31.781 | 23 | .105* | 68.188 | 22 | .000 | 63.315 | 20 | .000 | 73.138 | 22 | .000 | 54.744 | 21 | .000 |
| 5 | 123.105 | 31 | .000 | 148.907 | 31 | .000 | 124.812 | 32 | .000 | 103.070 | 31 | .000 | 142.404 | 32 | .000 |
| 6 | 94.887 | 27 | .000 | 87.247 | 29 | .000 | 87.090 | 27 | .000 | 70.378 | 29 | .000 | 106.318 | 29 | .000 |
| 7 | 60.867 | 27 | .000 | 82.625 | 27 | .000 | 74.545 | 28 | .000 | 49.320 | 27 | .006 | 64.592 | 29 | .000 |
| 8 | 37.707 | 31 | .189* | 46.055 | 31 | .040 | 38.393 | 32 | .202* | 48.461 | 31 | .024 | 58.481 | 33 | .004 |
| 9 | 68.500 | 30 | .000 | 87.038 | 30 | .000 | 67.078 | 31 | .000 | 55.032 | 29 | .003 | 74.055 | 30 | .000 |
| 10 | 42.684 | 34 | .146* | 47.717 | 33 | .047 | 48.075 | 35 | .069* | 53.739 | 33 | .013 | 50.592 | 33 | .026 |
| 11 | 54.656 | 28 | .002 | 50.938 | 29 | .007 | 53.252 | 30 | .006 | 82.395 | 29 | .000 | 65.346 | 30 | .000 |
| 12 | 44.052 | 31 | .060 | 62.688 | 30 | .000 | 68.252 | 33 | .000 | 48.784 | 30 | .017 | 85.960 | 32 | .000 |
| 13 | 61.602 | 27 | .000 | 53.651 | 27 | .002 | 47.459 | 28 | .012 | 55.658 | 28 | .001 | 60.588 | 29 | .001 |
| Total | 808.598 | 357 | .000 | 888.547 | 352 | .000 | 840.405 | 361 | .000 | 838.302 | 354 | .000 | 916.175 | 365 | .000 |

*Note*: * denotes observed and expected frequencies are not statistically different (α > .01)

Just like in the PCM and GPCM analyses, within the GGUM analyses, 10 fit groups were specified for the calculation of item and scale level model fit statistics. The GGUM2004 software collapses cells if the expected value of any response category for a group is zero. Collapsing of cells occurs separately at the item level. According to all 10 GGUM analyses, although the GGUM model fit statistically well for items 1, 3, and 4, the number of fit groups used and thus the degrees of freedom were very low for these fit analyses. Consideration of this point should be made when interpreting these results. The GGUM appeared to fit well for item 2 in half of the samples, though the same caution should be noted as those for items 1, 3, and 4. Finally, item 8 fit well in 4 of the 10 the samples. Overall, according to the results in Table 4.15, the goodness of fit of the GGUM for the Empowerment scale across 10 samples was not good.

Model fit was also examined for the Empowerment scale as a whole across the PCM, GPCM, and GGUM models by calculating AIC and BIC fit criteria. Table 4.16 displays those fit statistics estimated by the PCM, GPCM, and GGUM models.

Table 4.15

*Item and Scale Level Chi-Square Fit Statistics for Each Empowerment Sample: GGUM*

| | Sample 1 | | | | Sample 2 | | | | Sample 3 | | | | Sample 4 | | | | Sample 5 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Item | $\chi^2$ | *df* | *p* | *fit grps.* | $\chi^2$ | *df* | *p* | *fit grps.* | $\chi^2$ | *df* | *p* | *fit grps.* | $\chi^2$ | *df* | *p* | *fit grps.* | $\chi^2$ | *df* | *p* | *fit grps.* |
| 1 | 1.304 | 4 | .861* | 1 | 0.899 | 4 | .925* | 1 | 1.033 | 4 | .905* | 1 | 5.833 | 4 | .212* | 1 | 3.026 | 4 | .553* | 1 |
| 2 | 32.455 | 12 | .001 | 3 | 34.008 | 16 | .005 | 4 | 40.597 | 16 | .001 | 4 | 17.663 | 8 | .024 | 2 | 37.111 | 16 | .002 | 4 |
| 3 | 1.519 | 4 | .823* | 1 | 1.403 | 4 | .844* | 1 | 1.019 | 4 | .907* | 1 | 13.437 | 8 | .098* | 2 | 1.849 | 4 | .763* | 1 |
| 4 | 5.742 | 8 | .676* | 2 | 9.439 | 8 | .307* | 2 | 1.826 | 8 | .986* | 2 | 19.849 | 8 | .011 | 2 | 1.861 | 4 | .761* | 1 |
| 5 | 61.358 | 36 | .005 | 9 | 87.276 | 36 | .000 | 9 | 72.888 | 36 | .000 | 9 | 63.724 | 32 | .001 | 8 | 92.634 | 36 | .000 | 9 |
| 6 | 43.25 | 20 | .002 | 5 | 51.872 | 24 | .001 | 6 | 60.668 | 24 | .000 | 6 | 68.42 | 24 | .000 | 6 | 90.325 | 24 | .000 | 6 |
| 7 | 50.221 | 24 | .001 | 6 | 41.325 | 24 | .015 | 6 | 49.663 | 28 | .007 | 7 | 58.779 | 24 | .000 | 6 | 63.763 | 24 | .000 | 6 |
| 8 | 54.2 | 36 | .026 | 9 | 35.303 | 36 | .501* | 9 | 53.028 | 36 | .033 | 9 | 51.999 | 36 | .041 | 9 | 62.455 | 36 | .004 | 9 |
| 9 | 20.621 | 20 | .419* | 5 | 40.453 | 20 | .004 | 5 | 33.989 | 20 | .026 | 5 | 49.728 | 20 | .000 | 5 | 52.721 | 20 | .000 | 5 |
| 10 | 24.629 | 12 | .017 | 3 | 23.881 | 12 | .021 | 3 | 29.053 | 16 | .024 | 4 | 33.249 | 12 | .001 | 3 | 16.18 | 12 | .183* | 3 |
| 11 | 36.288 | 20 | .014 | 5 | 48.367 | 20 | .000 | 5 | 51.969 | 20 | .000 | 5 | 55.819 | 20 | .000 | 5 | 57.821 | 24 | .000 | 6 |
| 12 | 52.31 | 16 | .000 | 4 | 39.338 | 16 | .001 | 4 | 29.269 | 12 | .004 | 3 | 37.949 | 16 | .002 | 4 | 25.706 | 12 | .012 | 3 |
| 13 | 38.03 | 20 | .009 | 5 | 39.305 | 20 | .006 | 5 | 70.672 | 24 | .000 | 6 | 35.56 | 20 | .017 | 5 | 40.372 | 24 | .020 | 6 |
| Total | 421.93 | 58 | .000 | | 452.87 | 60 | .000 | | 495.67 | 62 | .000 | | 512.01 | 58 | .000 | | 545.82 | 60 | .000 | |

*Note*: * denotes observed and expected frequencies are not statistically different (α > .01)

Table 4.15 Con’t

*Item and Scale Level Chi-Square Fit Statistics for Each Empowerment Sample: GGUM*

| | Sample 6 | | | | Sample 7 | | | | Sample 8 | | | | Sample 9 | | | | Sample 10 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Item | $\chi^2$ | *df* | *p* | *fit grps.* | $\chi^2$ | *df* | *p* | *fit grps.* | $\chi^2$ | *df* | *p* | *fit grps.* | $\chi^2$ | *df* | *p* | *fit grps.* | $\chi^2$ | *df* | *p* | *fit grps.* |
| 1 | 1.745 | 4 | .782* | 1 | 2.096 | 4 | .718* | 1 | 0.791 | 4 | .939* | 1 | 1.272 | 4 | .866* | 1 | 2.1 | 4 | .717* | 1 |
| 2 | 40.305 | 16 | .001 | 4 | 3.083 | 8 | .929* | 2 | 6.161 | 12 | .908* | 3 | 22.167 | 16 | .138* | 4 | 13.872 | 12 | .309* | 3 |
| 3 | 10.986 | 8 | .202* | 2 | 1.582 | 4 | .812* | 1 | 0.909 | 4 | .923* | 1 | 1.32 | 4 | .858* | 1 | 2.995 | 4 | .559* | 1 |
| 4 | 2.196 | 8 | .974* | 2 | 1.455 | 4 | .834* | 1 | 9.155 | 8 | .329* | 2 | 12.699 | 8 | .123* | 2 | 2.015 | 4 | .733* | 1 |
| 5 | 88.33 | 36 | .000 | 9 | 94.721 | 36 | .000 | 9 | 94.936 | 36 | .000 | 9 | 89.621 | 36 | .000 | 9 | 114.76 | 36 | .000 | 9 |
| 6 | 62.709 | 24 | .000 | 6 | 53.933 | 24 | .000 | 6 | 58.79 | 24 | .000 | 6 | 57.658 | 24 | .000 | 6 | 89.208 | 24 | .000 | 6 |
| 7 | 41.759 | 24 | .014 | 6 | 95.203 | 28 | .000 | 7 | 47.125 | 24 | .003 | 6 | 38.157 | 24 | .033 | 6 | 65.852 | 28 | .000 | 7 |
| 8 | 25.924 | 36 | .893* | 9 | 45.443 | 36 | .134* | 9 | 56.336 | 36 | .017 | 9 | 50.231 | 36 | .058* | 9 | 60.315 | 36 | .007 | 9 |
| 9 | 35.357 | 20 | .018 | 5 | 49.121 | 20 | .000 | 5 | 43.977 | 20 | .002 | 5 | 49.22 | 20 | .000 | 5 | 46.687 | 20 | .001 | 5 |
| 10 | 20.428 | 16 | .201* | 4 | 14.989 | 12 | .242* | 3 | 19.431 | 12 | .079* | 3 | 26.392 | 12 | .009 | 3 | 20.972 | 12 | .051* | 3 |
| 11 | 40.063 | 20 | .005 | 5 | 63.014 | 20 | .000 | 5 | 37.671 | 20 | .010 | 5 | 77.418 | 20 | .000 | 5 | 59.334 | 20 | .000 | 5 |
| 12 | 28.298 | 12 | .005 | 3 | 8.319 | 8 | .403* | 2 | 55.746 | 12 | .000 | 3 | 33.717 | 12 | .001 | 3 | 37.147 | 12 | .000 | 3 |
| 13 | 62.401 | 24 | .000 | 6 | 36.465 | 20 | .014 | 5 | 31.614 | 20 | .048 | 5 | 43.584 | 20 | .002 | 5 | 44.826 | 24 | .006 | 6 |
| Total | 460.5 | 62 | .000 | | 469.42 | 56 | .000 | | 462.64 | 58 | .000 | | 503.46 | 59 | .000 | | 560.08 | 59 | .000 | |

*Note*: * denotes observed and expected frequencies are not statistically different (α > .01)

Table 4.16

*AIC and BIC criteria for each Empowerment Sample from PCM, GPCM, and GGUM models*

| Sample | PCM | | GPCM | | GGUM | |
|---|---|---|---|---|---|---|
| | AIC | BIC | AIC | BIC | AIC | BIC |
| 1 | 62220.041 | 62584.1 | 61307.645 | 61744.515 | 34865.542 | 35895.127 |
| 2 | 62266.819 | 62136.819 | 61365.618 | 61802.488 | 34946.166 | 35975.282 |
| 3 | 62188.445 | 62058.445 | 61310.701 | 61747.571 | 34903.493 | 35932.922 |
| 4 | 62256.564 | 62126.564 | 61416.119 | 61852.989 | 34948.076 | 35977.661 |
| 5 | 62342.067 | 62212.067 | 61415.376 | 61852.246 | 34954.672 | 35984.256 |
| 6 | 62875.969 | 62745.969 | 62026.278 | 62463.148 | 35574.157 | 36603.585 |
| 7 | 61769.049 | 61639.049 | 60793.885 | 61230.755 | 34350.537 | 35380.200 |
| 8 | 62003.826 | 61873.826 | 61060.273 | 61497.143 | 34615.964 | 35645.392 |
| 9 | 62213.792 | 62083.792 | 61368.296 | 61805.166 | 34927.587 | 35957.249 |
| 10 | 62342.181 | 62212.181 | 61351.74 | 61788.61 | 34955.237 | 35984.352 |

The AIC and BIC criteria are calculated in the same manner across the three IRT models. AIC and BIC statistics were also calculated for the CFA models, however, the factor analytic and IRT models, and the methods used to calculate AIC, are too discrepant to directly compare the AIC values, and therefore are not presented for the CFA analyses. The discrepancy lies in the fact that IRT models are models for response probabilities and factor analytic models are models for covariance and correlation matrices. The AIC and BIC criteria are directly comparable for the GPCM and GGUM analyses as both employed maximum likelihood methods for estimation, no priors were imposed on items, and the same prior (i.e., normal) was assumed for theta. However, a prior distribution of the discrimination (i.e., slope) parameter was necessary for the PCM analyses, where a log-normal prior distribution was specified in all PCM analyses. No priors were imposed for item parameters in the GGUM analyses. As a result, AIC and BIC criteria would be more consistent between the PCM and GGUM if Bayes estimation was employed.

Based on the results on Table 4.16, the GGUM analyses were consistently associated with the smallest criteria values, implying superior fit over the cumulative models. As for the cumulative models, the GPCM fit better than the PCM as evidenced by the smaller AIC and BIC values. Because no significance test exists for these statistics, they are to be interpreted as measures of relative differences between model outcomes. Therefore, the GGUM appeared to fit relatively much better than the GPCM, though the superiority of the GPCM over the PCM is not as prominent. In summary, analyses conducted to answer the fourth research question reveal that, across the three IRT models, the items generally do not fit well and neither do the models according to the chi-square statistics. According to the AIC and BIC criteria, the GGUM fit relatively better than the GPCM, though no criterion exists to measure ‘how much’ better.

### *Person Locations*

The focus of the second research question had to do with the location of the sample on the latent trait and the ordering of respondents on the latent trait across IRT models. IRT calibrations were conducted for the PCM, GPCM, and GGUM on the 10 Empowerment samples and person parameters are provided in this section. Rank-order correlations and scatterplots of the person parameters are presented to address research question two. Similar to the item location estimates, the theta estimates produced by the unfolding and cumulative IRT models are not analogous, and therefore comparisons cannot be direct and absolute. However, examination of correlations and distributions are appropriate. Table 4.17 displays

Kendall's Tau-*b* correlations of the person trait estimates across the 10 Empowerment samples for each pair of scaling methods. Correlations are presented for each sample, as a single mean may not capture small changes in rank ordering. This point is illustrated in Table 4.17 and the figures of scatterplots that follow. These correlations revealed that the PCM showed lower correlations with other scaling methods and the GGUM generally displayed higher correlations.

The high Tau-*b* correlation between the GPCM and GGUM models indicates that the rank ordering of respondents is essentially the same between models. Specifically regarding the GGUM, the lowest average Tau-*b* correlation was with the PCM (Tau-*b* = .888), indicating some inconsistency in terms of the rank ordering of people. However, all correlations in each of the 10 samples, in any scaling method combination were statistically significant ($p < .01$). Scatterplots in Figures 4.41, 4.42, and 4.43 depict the correspondence of trait estimates between the GGUM and the PCM, GPCM, and CFA models, respectively, from sample 1.

The majority of the cases within the first sample fell within the 95% confidence ellipse for the GGUM and PCM trait correlations indicating a fairly strong relationship between the two scaling methods. The GGUM and GPCM trait estimates are almost entirely coincident suggesting that the two models are essentially identical in scoring individuals. Although the Tau-*b* correlation between the GGUM and CFA models was relatively high, in the first sample (Tau = .957), the scatterplot in Figure 4.43 shows a slight non-linear relationship between the trait estimates and disagreement between the models towards the middle of the distribution.

Table 4.17

*Kendall's Tau-b Correlations among Person Trait Estimates by Sample and Scaling Method*

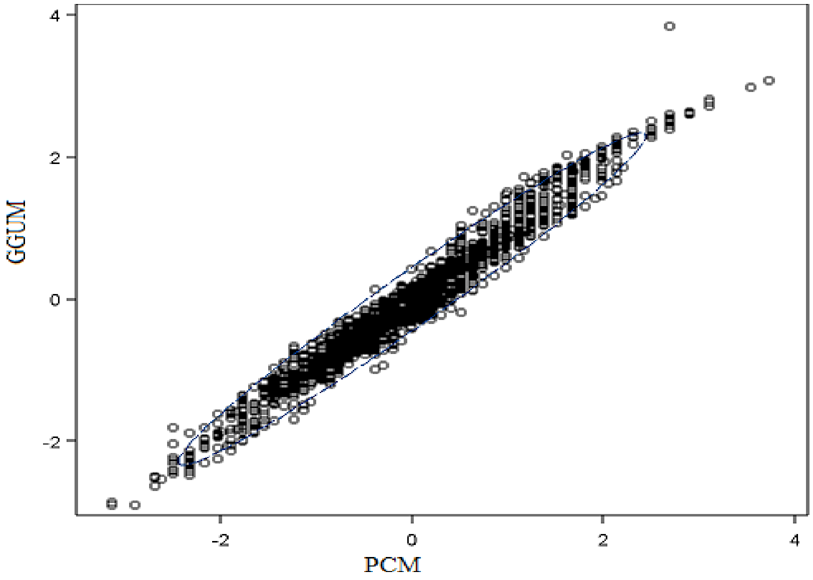
| Sample | PCM,<br>GGUM | GPCM,<br>GGUM | PCM,<br>GPCM | CFA,<br>PCM | CFA,<br>GPCM | CFA,<br>GGUM |
|---|---|---|---|---|---|---|
| 1 | .891 | .998 | .892 | .879 | .957 | .957 |
| 2 | .889 | .998 | .891 | .874 | .960 | .959 |
| 3 | .884 | .998 | .884 | .873 | .957 | .957 |
| 4 | .892 | .996 | .892 | .875 | .952 | .951 |
| 5 | .892 | .997 | .892 | .876 | .958 | .957 |
| 6 | .880 | .998 | .881 | .870 | .956 | .955 |
| 7 | .887 | .998 | .888 | .869 | .962 | .962 |
| 8 | .886 | .998 | .887 | .873 | .950 | .950 |
| 9 | .900 | .998 | .900 | .878 | .950 | .950 |
| 10 | .878 | .998 | .878 | .865 | .958 | .959 |
| Mean | .888 | .998 | .888 | .873 | .956 | .956 |

The relationship between the GGUM and CFA trait estimates is not entirely surprising given the assumption of a linear relationship between item responses and the latent trait that underlies the CFA model and the non-linearity that exists between probability of item endorsement and the latent trait within IRT models. The slight nonlinearity near the center of the each latent trait shows the CFA results yield higher trait values than the GGUM.

A closer examination of the trait distribution was facilitated by making the continuous trait distribution discrete by partitioning the trait distribution into quintiles and using 5 X 5 cross tabulation tables and using statistical measures of association appropriate for ordinal data. Results presented here correspond to the trait (i.e., person) estimates previously reported, in that cross tabulations for sample 1 are presented.  The overlap in frequencies of respondents within each quintile between the PCM, GPCM, and CFA methods with the GGUM are presented in Tables 4.18, Table 4.19, and Table 4.20.

Figure 4.41

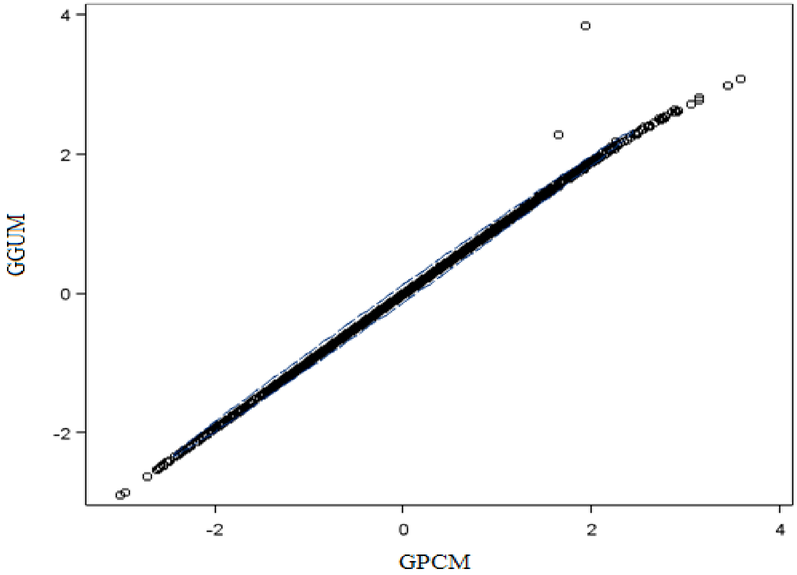
*Scatterplot of Trait Estimates for PCM and GGUM models: Sample 1 Empowerment Scale*



*Note: Dashed line denotes 95% prediction confidence ellipse*

Figure 4.42

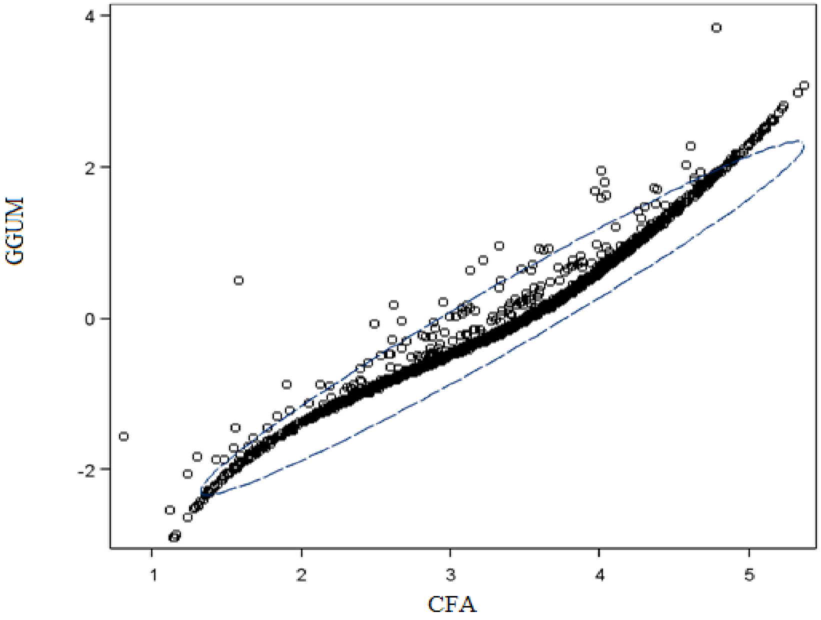
*Scatterplot of Trait Estimates for GPCM and GGUM models: Sample 1 Empowerment Scale*



*Note. Dashed line denotes 95% prediction confidence ellipse*

Figure 4.43

*Scatterplot of Trait Estimates for CFA and GGUM models: Sample 1 Empowerment Scale*



*Note, Dashed line denotes 95% prediction confidence ellipse*

The statistics presented in Tables 4.18, Table 4.19 and Table 4.20 are measures of association between the respective pairs of scaling methods. Kendall's Tau-*b* is a correlation that corrects for ties among data points, Stuart's Tau-c also accommodates ties and adjusts for the size of the table. Additionally, the Pearson correlation coefficient and the nonparametric Spearman rank correlation coefficient are tabled. These tables depict the discrepancies in the trait distribution across the pairs of scaling methods and tabled information of the graphical representations in the scatterplots. For example, Table 4.18 shows some dispersion of the trait estimates between the GGUM and PCM in all quintiles. Specifically, 43 people fell into the second PCM quintile, though were found to fall into the first quintile according to the GGUM distribution. Likewise, 42 trait estimates, according to the PCM distribution were located in the first quintile, while those 42 were located in the second GGUM quintile. Table 4.19 depicts the nearly identical trait distribution between the GGUM and GPCM, also as seen in Figure 4.42. The slightly higher trait values generated by the CFA over the GGUM seen towards the center of the distribution in the scatterplot (Figure 4.43) are more specifically differentiated in the cross-tabulation shown in Table 4.20. Additionally, there were nine observations where the GGUM yielded larger trait values than the CFA (i.e., two estimates fell within the 3rd GGUM quintile and the 1st CFA quintile).

Table 4.18

*Cross Tabulation Table of GGUM and PCM Quintiles: Sample 1, Empowerment Scale*

| | | PCM | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | Total | Statistic | Value | ASE |
| GGUM | 1 | 356 | 43 | 0 | 0 | 0 | 399 | Tau-*b* | .905 | .004 |
| | 2 | 42 | 296 | 62 | 0 | 0 | 400 | Tau-*c* | .905 | .004 |
| | 3 | 0 | 59 | 285 | 56 | 0 | 400 | Pearson | .950 | .003 |
| | 4 | 0 | 0 | 73 | 294 | 33 | 400 | Spearman | .950 | .003 |
| | 5 | 0 | 0 | 0 | 32 | 367 | 399 | | | |
| | Total | 398 | 398 | 420 | 382 | 400 | 1998 | | | |

*Note:* ASE = Asymptotic Standard Error; Tau-*b* = Kendall's Tau-*b*; Tau-*c* = Stuart's Tau-*c*

Table 4.19

*Cross Tabulation Table of GGUM and GPCM Quintiles: Sample 1, Empowerment Scale*

| | | | | GPCM | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | Total | Statistic | Value | ASE |
| GGUM | 1 | 398 | 1 | 0 | 0 | 0 | 399 | Tau-*b* | .999 | .001 |
| | 2 | 0 | 399 | 1 | 0 | 0 | 400 | Tau-*c* | .999 | .001 |
| | 3 | 0 | 0 | 399 | 1 | 0 | 400 | Pearson | .999 | .000 |
| | 4 | 0 | 0 | 0 | 399 | 1 | 400 | Spearman | .999 | .000 |
| | 5 | 0 | 0 | 0 | 0 | 399 | 399 | | | |
| | Total | 398 | 400 | 400 | 400 | 400 | 1998 | | | |

*Note.* ASE = Asymptotic Standard Error; Tau-*b* = Kendall's Tau-*b*; Tau-*c* = Stuart's Tau-*c*

Table 4.20

*Cross Tabulation Table of GGUM and CFA Quintiles: Sample 1, Empowerment Scale*

| | | | | CFA | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | Total | Statistic | Value | ASE |
| GGUM | 1 | 388 | 11 | 0 | 0 | 0 | 399 | Tau-*b* | .957 | .004 |
| | 2 | 7 | 366 | 27 | 0 | 0 | 400 | Tau-*c* | .957 | .004 |
| | 3 | 2 | 21 | 349 | 28 | 0 | 400 | Pearson | .977 | .002 |
| | 4 | 1 | 2 | 20 | 358 | 19 | 400 | Spearman | .976 | .002 |
| | 5 | 0 | 0 | 4 | 14 | 381 | 399 | | | |
| | Total | 398 | 400 | 400 | 400 | 400 | 1998 | | | |

*Note.* ASE = Asymptotic Standard Error; Tau-*b* = Kendall's Tau-*b*; Tau-*c* = Stuart's Tau-*c*

*Summary of Empowerment Analyses*

Although determination of dimensionality includes some amount of subjectivity, there was more supporting evidence that the assumptions of the cumulative models were met, compared to the unfolding models. Specifically, the data do not appear to be of the unfolding type based on the structure of the data in that two components did not emerge. Additionally the plot of pattern coefficients that resulted from the PCA with two components did not produce a circumplex-like (i.e., semi-circular) structure. Although not all fit indices produced by the CFA were supportive of excellent model fit, most indices did indicate that a single factor structure was adequate in explaining the empowerment data. This supports the assumptions of cumulative IRT models and the CFA method of scaling.

As for the item parameters, the PCM and GPCM performed similarly across all item parameters estimated. Both models indicated that most of the Empowerment items were easy to endorse, and that there are gaps on the latent trait that are not being measured by these 13 items. The GGUM analyses estimated all Empowerment items to have extreme location estimates and very large standard errors. According to the GGUM analyses, all items were clustered in an extreme region of the latent trait scale. The items themselves do not appear to be extremely worded in either direction, therefore the extremity of item parameter estimates could be an outcome of relative homogeneity of respondents' attitudes. Because item location estimates and signs of those estimates are also associated with item content, according to the GGUM analyses, moderate and negative attitudes towards teacher Empowerment are not well-measured by these 13 items. The category probability plots produced from the GGUM for the first four items show unfolding properties for the Strongly Agree and Agree response categories. The category probability plots produced from the PCM, GPCM, and GGUM for the remainder of the items are generally very similar, and imply that the three IRT models are functioning similarly.

Per the chi-square item level fit statistics, the PCM had few items that fit well, with the GPCM showing only slightly more statistically well-fitting items. At the scale level, both the PCM and GPCM models appeared to be statistically ill-fitting. The results from the GGUM item level chi-square fit statistics are inconclusive because not only must they be interpreted with caution, but the results are only indicators of gross item misfit, and not absolute statistical fit. Although the GGUM appeared to be the model that fit most Empowerment items, these statistically significant results could be a function of the small number of fit groups. At the scale level, the GGUM also did not fit well across the 10 samples. The final measures of fit, the AIC and BIC, indicated that the GGUM fit better than the GPCM, and that the GPCM fit better than the PCM.

Finally, the least amount of agreement between the person trait parameter estimates was found between the CFA and PCM scaling methods. The model that exhibited the least amount of agreement with the GGUM was the PCM, while the GPCM and GGUM were basically identical in estimating person parameters based on the rank order correlations and scatterplots.

## Leadership Scale

### *Testing Dimensionality Assumptions for Cumulative Models*

The first part of the fourth research question had to do with model assumptions and model fit. In practice testing model assumptions should precede analyses. To test model assumptions, such as dimensionality and independence, various procedures were conducted for all four scaling methods and are reported in this section for the Empowerment data. The same procedures used for analysis of the Empowerment scale data were used on the

Leadership data to assess model assumptions. A single factor model was specified for the entire sample using the 21 items that measured the construct teachers' perceptions of the leadership in their school. For model identification purposes, the factor loading was fixed at 1.0 for the item that possessed a large measure of variation. That item reads: "Overall, the school leadership in my school is effective."

Results from the single factor confirmatory analysis using the 21 Leadership items are reported in Table 4.21. The statistically significant (*p* < .05) model chi-square statistic and the RMSEA indicate that a single factor model does not fit well, however the RMR, SRMR, and NFI indicate reasonably good fit. The GFI is lower than 1.0, though with a value of .751 fit could be considered moderate. The large chi-square value is at least partly attributable to the large sample size. Inter-item correlations are helpful to consider in that high correlations could also contribute to a relatively high chi-square value. The inter-item correlation matrix of the 21 Leadership items contained correlations that were moderately high, with the majority of correlations ranging between .3 and .7. Overall, the single factor model does not appear to fit the Leadership data very well.

Table 4.21

*Fit Indices for the One Factor Leadership Model: Full Sample (n = 65,031)*

| Model $\chi^2$ | df | $\chi^2$/df | RMSEA | RMR | SRMR | NFI | GFI | Model AIC |
|---|---|---|---|---|---|---|---|---|
| 164815.747* | 189 | 872.041 | .135 | .056 | .056 | .954 | .751 | 205850.205 |

*Notes:* RMSEA = Root Mean Square Error of Approximation; RMR = Root Mean Square Residual; SRMR = Standardized Root Mean Square Residual, NFI = Normed Fit Index; GFI = Goodness of Fit Index, Model AIC = Akaike Information Criterion
* p < .05

*Local Independence*

Just as in the Empowerment analyses, the RMSEA for each item was calculated as a measure of relative fit at the item level. Local independence is closely related to the assumption of unidimensionality, and methods of assessing both overlap with measures of model fit. Table 4.22 displays the root mean square residuals at the item level from both the PCM and GPCM models where smaller values are interpreted as better relative fit. Based on these results, the smallest item residuals were associated with the GPCM, compared to the PCM model. The same statistic (i.e., chi-square item level likelihood-ratio fit statistics) for assessing local item independence was calculated for the GGUM model as was used in the Empowerment analyses. To summarize, items fit well in most of the GGUM analyses, although the fit statistics and associated *p*-values must be interpreted with caution. At the scale level, the GGUM did not fit statistically well in any analysis.

Table 4.22

*Item Level Residuals from PCM and GPCM Models: Leadership Scale*

| Item | PCM | GPCM | Item | PCM | GPCM |
|---|---|---|---|---|---|
| 1 | 5.056 | 4.900 | 12 | 9.062 | 7.759 |
| 2 | 9.695 | 9.112 | 13 | 8.297 | 7.579 |
| 3 | 9.781 | 5.548 | 14 | 7.810 | 6.637 |
| 4 | 5.552 | 5.818 | 15 | 4.979 | 5.760 |
| 5 | 6.912 | 6.655 | 16 | 6.896 | 6.078 |
| 6 | 8.217 | 7.842 | 17 | 5.042 | 5.965 |
| 7 | 6.139 | 7.262 | 18 | 10.605 | 5.289 |
| 8 | 12.957 | 4.807 | 19 | 12.055 | 5.429 |
| 9 | 8.775 | 7.626 | 20 | 5.305 | 6.870 |
| 10 | 8.550 | 5.533 | 21 | 14.954 | 22.274 |
| 11 | 27.996 | 11.193 | | | |

*Unidimensionality under Unfolding Models*

Assessing the dimensionality of the data from the ideal point response perspective, one principal components analysis was conducted on the Leadership data using the entire sample (*n* = 65,031) to assess dimensionality. Dimensionality of the Leadership data included the examination of eigenvalues, final communality estimates, pattern coefficients, and plots of pattern coefficients resulting from the application of principal components analyses. Generally, if an item level communality, generated from the first two components, is ≥ .3, then that item is not likely violating the assumption of unidimensionality (Roberts et al., 2000). Item level final communality estimates derived from a two factor component model for the 21 Leadership items are found in Table 4.23. Communalities for all items comprising were greater than .3.

Table 4.23

*Final Community Estimates for the Leadership Items (i = 21)*

| Item | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0.593 | 0.39 | 0.645 | 0.589 | 0.615 | 0.642 | 0.392 | 0.726 | 0.596 | 0.623 | 0.573 |
| Item | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | |
| | 0.775 | 0.762 | 0.718 | 0.565 | 0.649 | 0.542 | 0.745 | 0.754 | 0.537 | 0.509 | |

If the data are of the unfolding type and unidimensional, the first two eigenvalues of the PCA should be larger than the remaining eigenvalues. However, a prescribed criterion does not exist as a measure for "large." The first five eigenvalues from the Leadership analysis were: 11.700, 1.238, 1.049, .910, and .652. The second eigenvalue was not substantially larger than the third eigenvalue. This is at least in part, supporting evidence that the assumption of unidimensionality within the context of cumulative IRT models is supported, as opposed to unidimensionality within the context of unfolding models.

Finally, the plot of pattern coefficients produced from the PCA with two factors (i.e., components) was examined to determine the structure of the Leadership data. A semi-circular pattern of the coefficients in a two-factor space would be evidence that two linear principal components explain the pattern of data. This is because within the context of unfolding IRT models, there are two linear principal components for each unfolding dimension. Reported in Table 4.24 are the pattern coefficients from the two component PCA with the Leadership samples and the plot of pattern coefficients for the 21 Leadership items is displayed in Figure 4.44. In Figure 4.44, two linear components are also not evidenced from the pattern coefficients, suggesting that the structure of the Leadership data does not conform to that of the unfolding type (i.e., the responses to the 21 Leadership items do not unfold). These results suggest that application of unfolding models may not be necessary.

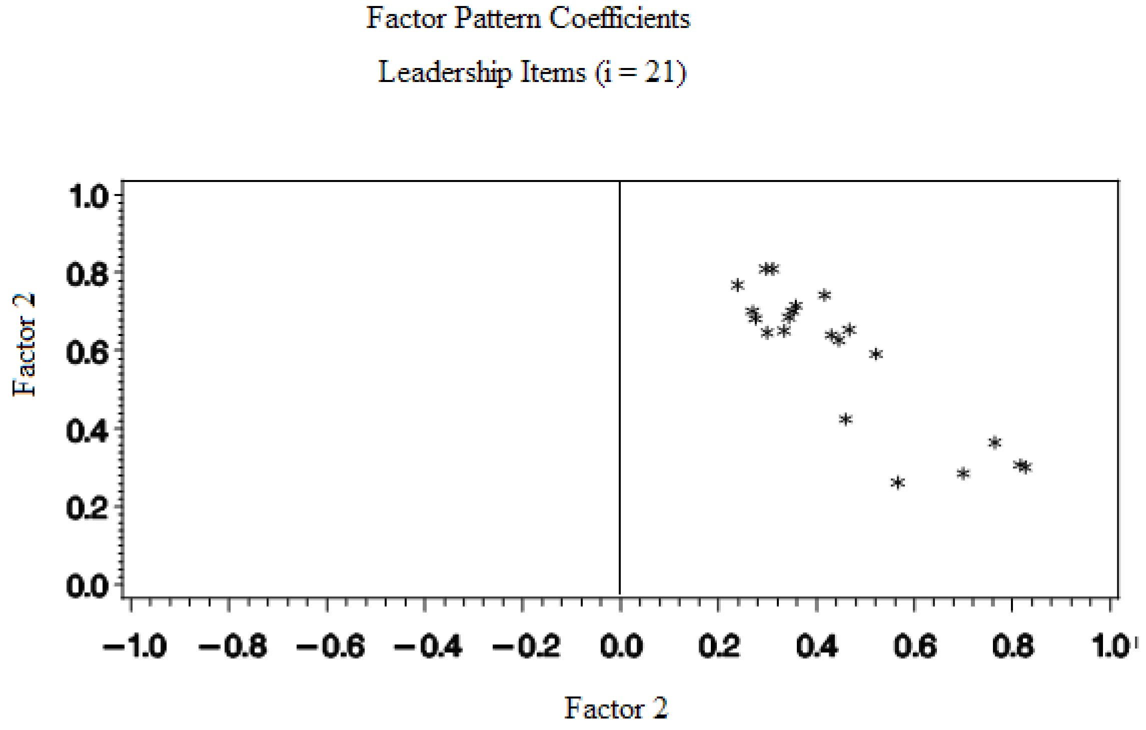
Table 4.24

*Factor Pattern Coefficients Derived From 2 Principal Components: Leadership*

| Items | Factor 1 | Factor 2 | Item | Factor 1 | Factor 2 |
|---|---|---|---|---|---|
| 1 | .627 | .447 | 12 | .302 | .827 |
| 2 | .264 | .566 | 13 | .308 | .817 |
| 3 | .654 | .467 | 14 | .365 | .764 |
| 4 | .686 | .344 | 15 | .701 | .270 |
| 5 | .701 | .352 | 16 | .769 | .239 |
| 6 | .716 | .358 | 17 | .683 | .275 |
| 7 | .425 | .459 | 18 | .810 | .297 |
| 8 | .743 | .417 | 19 | .810 | .312 |
| 9 | .641 | .431 | 20 | .652 | .334 |
| 10 | .592 | .522 | 21 | .647 | .300 |
| 11 | .286 | .700 | | | |

Figure 4.44

*Plot of Factor Pattern Coefficients for the Leadership Scale*



Development of the CFA Scale

The factor analytic scoring method and procedures used to create the Empowerment scale were mirrored for the Leadership data to transform the latent trait, Leadership, into an observed measure.  It was assumed that a single latent trait was measured with the 21 leadership items. Across the 21 items, the item that consistently displayed large measures of variation was: “Overall, the school leadership in my school is effective.” The standard deviation for this item ranged from 1.177 to 1.207, and the item’s factor loading was fixed at 1 across all 10 samples. The factor scores resulting from the CFA analyses were used to weight item responses, and then those products were summed to create the observed variable, Leadership for each respondent for all 10 samples. Various measures of model/data fit for the

10 leadership samples are shown in table 4.25. Although the RMSEA values are slightly high, and GFI indices are moderately low, overall these fit statistics indicate moderately good fit of the CFA model.

The preceding results indicate that, for most of the fit statistics, a single factor model fit moderately well; therefore the assumption of unidimensionality within the context of cumulative IRT models was likely not violated. Further, it can be concluded that the 21 items measured the construct, Leadership in the CFA model reasonably well also. Unidimensionality within the context of unfolding IRT models was not met given that two components did not explain the data, and that the plots of the pattern coefficients did not form a semi-circular shape.

Table 4.25

*Fit Indices for the One Factor Leadership Model by Sample (n = 10)*

| Sample | Model $\chi^2$ | df | $\chi^2$/df | RMSEA | RMR | SRMR | NFI | GFI | Model AIC |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 5632.244 | 189 | 29.800 | .138 | .058 | .058 | .948 | .739 | 6861.829 |
| 2 | 5556.350 | 189 | 29.399 | .140 | .059 | .059 | .951 | .733 | 7151.897 |
| 3 | 5605.735 | 189 | 29.660 | .138 | .061 | .061 | .945 | .738 | 6857.273 |
| 4 | 5701.966 | 189 | 30.169 | .140 | .060 | .060 | .947 | .734 | 7085.787 |
| 5 | 5356.682 | 189 | 28.342 | .135 | .058 | .058 | .952 | .746 | 6556.885 |
| 6 | 5268.387 | 189 | 27.875 | .134 | .055 | .055 | .951 | .750 | 6451.392 |
| 7 | 5471.779 | 189 | 28.951 | .135 | .058 | .058 | .949 | .746 | 6574.815 |
| 8 | 5462.469 | 189 | 28.902 | .138 | .061 | .061 | .948 | .738 | 6878.166 |
| 9 | 5392.454 | 189 | 28.532 | .136 | .056 | .056 | .951 | .743 | 6690.514 |
| 10 | 5061.434 | 189 | 26.780 | .131 | .054 | .054 | .956 | .757 | 6298.652 |

*Notes:* RMSEA = Root Mean Square Error of Approximation; RMR = Root Mean Square Residual; SRMR = Standardized Root Mean Square Residual, NFI = Normed Fit Index; GFI = Goodness of Fit Index, Model AIC = Akaike Information Criterion

## IRT Parameter Estimates

### *Item Locations*

The second research question focused on the item location estimates generated from the three IRT models, as it was hypothesized that locations may be very different across the two types of IRT models (cumulative, unfolding), if the survey was constructed using a method that assumed a dominance response process. To investigate this, IRT calibrations were performed on all 10 Leadership data sets with the application of the PCM, GPCM, and GGUM models. Item parameter estimates, including the location estimates, as presented in this section.

The Leadership items as they appear on the NCTWC survey are presented in Table 4.26. Presented in Table 4.27 are the item location estimates averaged across the 10 samples, with the average standard errors by IRT model, and the order in which each model ranks the average item location estimates. As in the Empowerment analyses, the underlying trait, Leadership, and the respective scale upon which both item and person estimates are located, are not directly comparable for the cumulative (PCM, GPCM) and GGUM models.

The parameter estimation for the PCM and GPCM results in Table 4.27 are not directly comparable to the GGUM estimates as theta and the resulting scale are different. All item location parameters generated from the 10 GGUM analyses on the Leadership items were located in the same general, and relatively extreme region of the latent trait ranging from $\delta_i$ = 3.314 (item 21) to $\delta_i$ = 4.041 (item 19). This clustering of items indicates that the full range of the latent trait (attitudes towards or about Leadership) is not well measured, only a narrow interval. As in the Empowerment analyses, items with extreme GGUM estimates also had the largest standard errors.

Table 4.26

*Leadership Items (i = 21)*

| Leadership |
|---|
| Please rate your level of agreement with the following statements: |
| 1. There is an atmosphere of trust and mutual respect within the school<br>2. The faculty are committed to helping every student learn<br>3. The school leadership communicates clear expectations to students and parents<br>4. The school leadership shields teachers from disruptions, allowing teachers to focus on educating students<br>5. The school leadership consistently enforces rules for student conduct<br>6. The school leadership support teachers' efforts to maintain discipline in the classroom<br>7. Opportunities are available for members of the community to actively contribute to this school's success<br>8. The school leadership consistently supports teachers |
| Please rate your level of agreement with the following statements: |
| 9. The school improvement team provides effective leadership at this school<br>10. The faculty and staff have a shared vision<br>11. Teachers are held to high professional standards for delivering instruction<br>12. Teacher performance evaluations are handled in an appropriate manner<br>13. The procedures for teacher performance evaluations are consistent<br>14. Teachers receive feedback that can help them improve teaching |
| The school leadership makes a sustained effort to address teacher concerns about: |
| 15. Facilities and resources<br>16. The use of time in my school<br>17. Professional development<br>18. Empowering teachers<br>19. Leadership issues<br>20. New teacher support<br>21. Overall, the school leadership in my school is effective |

Table 4.27

*Average Item Location, Standard Errors, and Rank Orders of Item Locations across 10 Samples: Leadership (i=21)*

| | Average Item Location<br>(Standard Error) | | | Rank Order of Average Item Locations | | |
|---|---|---|---|---|---|---|
| Item | PCM | GPCM | GGUM | PCM | GPCM | GGUM |
| 1 | -0.376 | -0.380 | 3.827 | 6 | 6 | 4 |
| | (.030) | (.031) | (3.108) | | | |
| 2 | -1.136 | -1.366 | 3.378 | 21 | 21 | 15 |
| | (.032) | (.043) | (2.025) | | | |
| 3 | -0.661 | -0.636 | 3.576 | 15 | 15 | 12 |
| | (.030) | (.029) | (7.814) | | | |
| 4 | -0.331 | -0.337 | 3.723 | 3 | 4 | 5 |
| | (.029) | (.030) | (2.914) | | | |
| 5 | -0.218 | -0.220 | 3.901 | 1 | 1 | 3 |
| | (.031) | (.032) | (3.862) | | | |
| 6 | -0.498 | -0.488 | 3.634 | 10 | 10 | 9 |
| | (.032) | (.031) | (5.391) | | | |
| 7 | -0.984 | -1.093 | 3.294 | 19 | 19 | 17 |
| | (.032) | (.037) | (1.714) | | | |
| 8 | -0.552 | -0.517 | 3.718 | 12 | 11 | 6 |
| | (.031) | (.028) | (33.504) | | | |
| 9 | -0.431 | -0.429 | 3.711 | 8 | 7 | 7 |
| | (.031) | (.030) | (3.449) | | | |
| 10 | -0.571 | -0.555 | 3.649 | 14 | 13 | 8 |
| | (.030) | (.030) | (5.195) | | | |
| 11 | -1.029 | -1.099 | 3.111 | 20 | 20 | 20 |
| | (.031) | (.036) | (2.016) | | | |
| 12 | -0.839 | -0.839 | 3.137 | 18 | 18 | 19 |
| | (.035) | (.039) | (2.587) | | | |
| 13 | -0.762 | -0.788 | 3.157 | 16 | 17 | 18 |
| | (.035) | (.039) | (2.425) | | | |
| 14 | -0.770 | -0.775 | 3.298 | 17 | 16 | 16 |
| | (.032) | (.034) | (3.058) | | | |
| 15 | -0.524 | -0.533 | 3.444 | 11 | 12 | 14 |
| | (.031) | (.033) | (2.466) | | | |
| 16 | -0.359 | -0.355 | 3.621 | 5 | 5 | 10 |
| | (.031) | (.031) | (3.540) | | | |
| 17 | -0.566 | -0.574 | 3.456 | 13 | 14 | 13 |
| | (.032) | (.033) | (2.523) | | | |
| 18 | -0.332 | -0.319 | 3.996 | 4 | 3 | 2 |
| | (.033) | (.029) | (26.151) | | | |
| 19 | -0.317 | -0.294 | 4.041 | 2 | 2 | 1 |
| | (.033) | (.029) | (42.119) | | | |
| 20 | -0.443 | -0.457 | 3.599 | 9 | 9 | 11 |
| | (.030) | (.032) | (2.538) | | | |
| 21 | -0.395 | -0.435 | 3.314 | 7 | 8 | 21 |
| | (.030) | (.038) | (1.940) | | | |

As in the Empowerment analyses, the PCM and GPCM functioned similarly in terms of the average item locations and the ranking of those estimates. In the PCM analyses, the average item locations ranged from $b$ = -1.136 (item 21) to $b$ = -.218 (item 1). Average item estimates for most of the 21 items were nearly identical between the two cumulative models. The location estimates reveal that all of the Leadership items were generally easy to endorse, with item 21 being the easiest, on average. According to the PCM and GPCM analyses, none of the items on the Leadership scale required a strong positive attitude towards Leadership.

The average correlation across the 10 samples between the PCM and GPCM for estimated item locations was .953. The average correlation between the PCM and GGUM was .611 and the correlation between GPCM and GGUM item estimates was .641. Across the 10 Leadership samples, all correlations between the PCM and GPCM location estimates were statistically significant ($p < .05$). All of the correlations between the PCM and GGUM, and between the GPCM and GGUM location estimates were statistically significant ($p < .05$). These correlations, associated $p$-values and comparable rank ordering of item locations indicate that the PCM and GPCM functioned similarly.

The discriminating characteristics of the Leadership items are described next for the three IRT models. Presented in Table 4.28 are the average item discriminations ($a$ parameters) and standard errors across the 10 samples. In the Leadership analyses, the item discrimination ($a$) parameter was fixed to have a mean of 1.0 and a standard deviation of .0001 for the PCM calibration, just as in the Empowerment analyses. The discrimination parameter is estimated in the GPCM models, and calibration converged with fewer E-M iterations than the PCM models.

Table 4.28

*Average Item Discrimination and Standard Errors across 10 Leadership Samples*

| | Average Item Discrimination | | | Average Standard Error | | |
|---|---|---|---|---|---|---|
| Item | PCM | GPCM | GGUM | PCM | GPCM | GGUM |
| 1 | 1.053 | 1.018 | 1.668 | 0.005 | 0.031 | 0.062 |
| 2 | 1.053 | 0.608 | 0.96 | 0.005 | 0.02 | 0.055 |
| 3 | 1.053 | 1.373 | 2.255 | 0.005 | 0.045 | 0.088 |
| 4 | 1.053 | 0.974 | 1.589 | 0.005 | 0.032 | 0.061 |
| 5 | 1.053 | 1.008 | 1.651 | 0.005 | 0.035 | 0.061 |
| 6 | 1.053 | 1.229 | 2.015 | 0.005 | 0.042 | 0.076 |
| 7 | 1.053 | 0.758 | 1.212 | 0.005 | 0.024 | 0.059 |
| 8 | 1.053 | 1.782 | 2.937 | 0.005 | 0.062 | 0.113 |
| 9 | 1.053 | 1.142 | 1.871 | 0.005 | 0.036 | 0.065 |
| 10 | 1.053 | 1.28 | 2.098 | 0.005 | 0.041 | 0.076 |
| 11 | 1.053 | 0.836 | 1.336 | 0.005 | 0.027 | 0.067 |
| 12 | 1.053 | 1.043 | 1.677 | 0.005 | 0.045 | 0.074 |
| 13 | 1.053 | 0.992 | 1.607 | 0.005 | 0.043 | 0.068 |
| 14 | 1.053 | 1.065 | 1.725 | 0.005 | 0.039 | 0.073 |
| 15 | 1.053 | 1.008 | 1.633 | 0.005 | 0.032 | 0.065 |
| 16 | 1.053 | 1.148 | 1.87 | 0.005 | 0.038 | 0.068 |
| 17 | 1.053 | 1.025 | 1.662 | 0.005 | 0.032 | 0.065 |
| 18 | 1.053 | 1.624 | 2.678 | 0.005 | 0.058 | 0.099 |
| 19 | 1.053 | 1.761 | 2.902 | 0.005 | 0.063 | 0.105 |
| 20 | 1.053 | 0.905 | 1.469 | 0.005 | 0.028 | 0.058 |
| 21 | 1.053 | 0.6 | 0.989 | 0.005 | 0.014 | 0.032 |

However, the Leadership PCM analyses across the 10 samples took approximately 80 iterations, while the GPCM generally required about 50 E-M iterations to converge. By comparison, in the Empowerment analyses, each PCM analysis took approximately 50 iterations of the E-M cycle to converge while the GPCM analyses required about 25 cycles. Although the PCM required more iterations overall, the Leadership data required more iteration regardless of IRT model, in part due to the increased number of item parameters to be estimated.

Just as in the Empowerment analyses, the GPCM and GGUM ordered the *a* parameters almost identically. Item 8 ("The school leadership consistently supports teachers") was the most discriminating, on average, and items 2 ("The faculty are committed to helping every student learn") and 21 ("Overall, the school leadership in my school is effective") exhibited the lowest discrimination parameters. The average Kendall's Tau-*b* correlation between the GPCM and GGUM discrimination parameter estimates, across the 10 Leadership samples was .899. These correlations, across the 10 samples were all statistically significant ($p < .05$).

Relative to the other 20 Leadership items, item 2 exhibited low discrimination due to the fact that just under 84% of the entire sample agreed or strongly agreed with this item. This item did not discriminate well among respondents, as most people agreed regardless of their standing on the latent trait. For example, items 8, 18, and 19 had relatively high discrimination parameter estimates, meaning that these items differentiated well between those respondents with low and high levels of the latent trait, attitude towards/about school Leadership. Although these three items had discrimination estimates that were higher relative to the other items (see Table 4.29), the majority of the sample either agreed or strongly agreed with all of the Leadership items.

The final item parameters estimated using the IRT models were the category probability thresholds. The four category threshold parameters, $d_k$ , averaged across the 10 samples for the PCM and GPCM models are given in Table 4.30. Item step parameters, $b_{jk}$ , were also calculated in the Leadership analyses for the PCM and GPCM models, which represent the point on the latent trait where two adjacent category probability functions intersect. These average item step parameters are presented in Table 4.31 for all three IRT models. Just as for the Empowerment data, category 1 in the Leadership analyses represents Strongly Disagree and category 5 represents Strongly Agree.

Table 4.30

*Average Category Threshold Parameters across 10 Leadership Samples*

| | PCM | | | | GPCM | | | |
|---|---|---|---|---|---|---|---|---|
| Item | $d_2$ | $d_3$ | $d_4$ | $d_5$ | $d_2$ | $d_3$ | $d_4$ | $d_5$ |
| 1 | 1.098 | 0.142 | 0.432 | -1.67 | 1.118 | 0.135 | 0.452 | -1.71 |
| 2 | 1.241 | 0.012 | 0.437 | -1.69 | 1.472 | -0.24 | 0.901 | -2.13 |
| 3 | 1.073 | 0.14 | 0.368 | -1.58 | 1.036 | 0.195 | 0.25 | -1.48 |
| 4 | 1.16 | 0.116 | 0.309 | -1.58 | 1.196 | 0.097 | 0.345 | -1.64 |
| 5 | 1.053 | 0.087 | 0.382 | -1.52 | 1.07 | 0.08 | 0.406 | -1.56 |
| 6 | 0.982 | 0.198 | 0.38 | -1.56 | 0.973 | 0.225 | 0.312 | -1.51 |
| 7 | 1.326 | 0.551 | 0.084 | -1.96 | 1.445 | 0.606 | 0.194 | -2.25 |
| 8 | 1.021 | 0.256 | 0.233 | -1.51 | 0.958 | 0.304 | 0.084 | -1.35 |
| 9 | 1.239 | 0.705 | -0.07 | -1.88 | 1.242 | 0.689 | -0.08 | -1.86 |
| 10 | 1.187 | 0.419 | 0.215 | -1.82 | 1.163 | 0.423 | 0.153 | -1.74 |
| 11 | 0.943 | 0.18 | 0.487 | -1.61 | 0.981 | 0.128 | 0.654 | -1.76 |
| 12 | 0.917 | 0.253 | 0.414 | -1.58 | 0.938 | 0.248 | 0.423 | -1.61 |
| 13 | 0.99 | 0.285 | 0.314 | -1.59 | 1.016 | 0.278 | 0.339 | -1.63 |
| 14 | 1.024 | 0.34 | 0.28 | -1.64 | 1.042 | 0.339 | 0.274 | -1.66 |
| 15 | 1.282 | 0.423 | 0.359 | -2.06 | 1.314 | 0.421 | 0.38 | -2.12 |
| 16 | 1.395 | 0.36 | 0.298 | -2.05 | 1.381 | 0.373 | 0.266 | -2.02 |
| 17 | 1.244 | 0.446 | 0.386 | -2.08 | 1.27 | 0.445 | 0.398 | -2.11 |
| 18 | 1.217 | 0.477 | 0.214 | -1.91 | 1.147 | 0.48 | 0.106 | -1.73 |
| 19 | 1.283 | 0.551 | 0.199 | -2.03 | 1.193 | 0.539 | 0.086 | -1.82 |
| 20 | 1.116 | 0.484 | 0.21 | -1.81 | 1.16 | 0.493 | 0.266 | -1.92 |
| 21 | 0.792 | 0.312 | 0.386 | -1.49 | 0.786 | 0.274 | 0.761 | -1.82 |

*Notes:* $d_2$ = threshold parameter for category 2 (Disagree); $d_3$ = threshold parameter for category 3 (Neither Agree Nor Disagree); $d_4$ = threshold parameter for category 4 (Agree); $d_5$ = threshold parameter for category 5 (Strongly Agree)

Table 4.31

*Average Category Step and Threshold Parameters Across 10 Leadership Samples*

| | PCM | | | | GPCM | | | | GGUM | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Item | $b_{j2}$ | $b_{j3}$ | $b_{j4}$ | $b_{j5}$ | $b_{j2}$ | $b_{j3}$ | $b_{j4}$ | $b_{j5}$ | $\tau_{i2}$ | $\tau_{i3}$ | $\tau_{i4}$ | $\tau_{i5}$ |
| 1 | -1.475 | -0.518 | -0.808 | 1.295 | -1.498 | -0.515 | -0.832 | 1.326 | -5.393 | -4.369 | -4.700 | -2.459 |
| 2 | -2.377 | -1.148 | -1.573 | 0.554 | -2.839 | -1.124 | -2.267 | 0.764 | -6.455 | -4.545 | -5.805 | -2.513 |
| 3 | -1.734 | -0.801 | -1.028 | 0.921 | -1.672 | -0.831 | -0.886 | 0.845 | -5.322 | -4.446 | -4.502 | -2.712 |
| 4 | -1.491 | -0.447 | -0.640 | 1.253 | -1.533 | -0.433 | -0.682 | 1.301 | -5.326 | -4.178 | -4.442 | -2.375 |
| 5 | -1.271 | -0.305 | -0.600 | 1.303 | -1.291 | -0.300 | -0.626 | 1.336 | -5.250 | -4.219 | -4.559 | -2.524 |
| 6 | -1.480 | -0.695 | -0.877 | 1.062 | -1.461 | -0.714 | -0.800 | 1.021 | -5.161 | -4.383 | -4.472 | -2.586 |
| 7 | -2.310 | -1.534 | -1.067 | 0.977 | -2.538 | -1.699 | -1.287 | 1.152 | -6.010 | -5.075 | -4.659 | -2.054 |
| 8 | -1.574 | -0.808 | -0.786 | 0.958 | -1.475 | -0.820 | -0.601 | 0.828 | -5.258 | -4.578 | -4.347 | -2.874 |
| 9 | -1.670 | -1.137 | -0.365 | 1.446 | -1.671 | -1.118 | -0.353 | 1.426 | -5.458 | -4.879 | -4.086 | -2.243 |
| 10 | -1.757 | -0.990 | -0.786 | 1.250 | -1.718 | -0.977 | -0.708 | 1.184 | -5.444 | -4.671 | -4.392 | -2.434 |
| 11 | -1.972 | -1.209 | -1.516 | 0.581 | -2.081 | -1.228 | -1.754 | 0.665 | -5.307 | -4.393 | -4.968 | -2.399 |
| 12 | -1.756 | -1.092 | -1.253 | 0.745 | -1.778 | -1.087 | -1.262 | 0.770 | -5.006 | -4.279 | -4.471 | -2.339 |
| 13 | -1.752 | -1.047 | -1.076 | 0.828 | -1.804 | -1.066 | -1.127 | 0.846 | -5.044 | -4.266 | -4.342 | -2.264 |
| 14 | -1.793 | -1.109 | -1.049 | 0.874 | -1.816 | -1.114 | -1.049 | 0.881 | -5.202 | -4.464 | -4.402 | -2.385 |
| 15 | -1.806 | -0.947 | -0.883 | 1.540 | -1.847 | -0.954 | -0.913 | 1.582 | -5.381 | -4.442 | -4.407 | -1.794 |
| 16 | -1.754 | -0.719 | -0.657 | 1.694 | -1.735 | -0.728 | -0.621 | 1.665 | -5.436 | -4.383 | -4.276 | -1.902 |
| 17 | -1.810 | -1.012 | -0.952 | 1.509 | -1.844 | -1.019 | -0.972 | 1.539 | -5.389 | -4.522 | -4.480 | -1.854 |
| 18 | -1.549 | -0.809 | -0.546 | 1.576 | -1.466 | -0.799 | -0.425 | 1.414 | -5.526 | -4.832 | -4.443 | -2.551 |
| 19 | -1.601 | -0.868 | -0.517 | 1.716 | -1.488 | -0.833 | -0.381 | 1.524 | -5.594 | -4.912 | -4.442 | -2.484 |
| 20 | -1.559 | -0.927 | -0.653 | 1.366 | -1.617 | -0.951 | -0.723 | 1.462 | -5.295 | -4.595 | -4.364 | -2.074 |
| 21 | -1.187 | -0.707 | -0.782 | 1.094 | -1.221 | -0.709 | -1.196 | 1.386 | -4.583 | -4.046 | -4.569 | -1.800 |

The average (across the 10 samples) points on the latent trait scale, where the category 1 (Strongly Disagree) and category 2 (Disagree) probabilities intersect within the PCM and GPCM models are given in the column labeled $b_{j2}$ in Table 4.31. The same interpretation is made for the points on the latent trait where the intersection of the rest of the adjacent probability functions intersect in the columns labeled $b_{j3}$, $b_{j4}$, and $b_{j5}$. For example, across the averaged parameters estimated within the PCM analyses for item 1, the point on the latent trait (Leadership) scale, where the probability of endorsing category 1 (Strongly Disagree) and category 2 (Disagree) intersect, is located on average, at -1.475.

The average step parameters were not evenly distributed across the latent trait, in that the middle two average step parameters generated by the PCM and GPCM ($b_{j3}$, $b_{j4}$) were estimated to be very close. Also, the point of intersection between strongly agree and agree ($b_{j4}$), across all 21 items was generally higher than the adjacent step parameter. This discrepancy between $b_{j5}$ and $b_{j4}$ in the Leadership analyses was, however, much less pronounced than that within the Empowerment analyses. These results support the observed responses by category in Table 4.29 that respondents did not use the five response categories evenly; a disproportionate percentage of the respondents agreed with all the items.

Table 4.29

*Percentage of Category Endorsement by Leadership Item: Full Sample (n = 65,012)*

| | Item | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Category | Strongly Disagree | 7.4 | 1.3 | 4.5 | 8 | 10.9 | 6.7 | 1.4 | 5.6 | 4.5 | 4 | 2.1 | 3.4 | 3.6 | 3.4 | 3.7 | 4.9 | 3.4 | 6 | 5.6 | 5.4 | 9.1 |
| | Disagree | 16.5 | 5.8 | 11.4 | 18.9 | 20.5 | 13.2 | 5.1 | 13 | 12 | 11.3 | 5.4 | 7.3 | 8.4 | 8.5 | 11.3 | 15.8 | 10.3 | 14.7 | 14.3 | 12.2 | 13 |
| | Neither Agree Nor Disagree | 14.1 | 8.1 | 12.8 | 15.3 | 13.8 | 14.3 | 16.8 | 15.6 | 27.4 | 18.5 | 8.5 | 11.3 | 13.5 | 14.2 | 16.2 | 18.7 | 16.4 | 20.9 | 22.4 | 19.9 | 14.9 |
| | Agree | 45.9 | 52.8 | 48.9 | 42.3 | 39.4 | 46.5 | 55.2 | 45 | 43 | 49.7 | 53.9 | 52.3 | 50.4 | 50.8 | 54.4 | 49 | 55.7 | 46 | 46.5 | 47.5 | 44.4 |
| | Strongly Agree | 15.4 | 31 | 21.9 | 15.3 | 14.9 | 18.9 | 21.1 | 20.2 | 12.6 | 15.9 | 29.9 | 25.5 | 23.8 | 22.6 | 12.3 | 10.5 | 13.3 | 11.6 | 10.4 | 14.1 | 18.3 |

Plots of the threshold parameters within the PCM, GPCM and GGUM analyses are useful for interpretation of the item parameter estimates previously described. Derivation and examination of the probability plots were the focus of the third research question, as it was hypothesized that, for the items that contained relatively neutral content, the plots would display characteristics of the ideal point response process (i.e., single-peaked, non-monotonic). It was also hypothesized that the two types of IRT models would function similarly if the attitudes possessed by the sample were located on one side of the items (i.e., homogeneous sample not measured well by items). Figures 4.45 through 4.65 display the category probability functions for the 21 Leadership items from application of the PCM on the first simple random sample. Figures 4.66 through 4.86 display the category probability functions for the 21 Leadership items from application of the GPCM on the first simple random sample, and Figures 4.87 through 4.107 display the category probability plots for the 21 Leadership items resulting from the GGUM analyses on the first sample. The category probability plots for the 21 Leadership items are displayed in Figures 4.66 through 4.86 resulting from the application of the GPCM.

Figure 4.45

*Category Probability Plot for Item 1 with PCM: Sample 1, Leadership Scale*

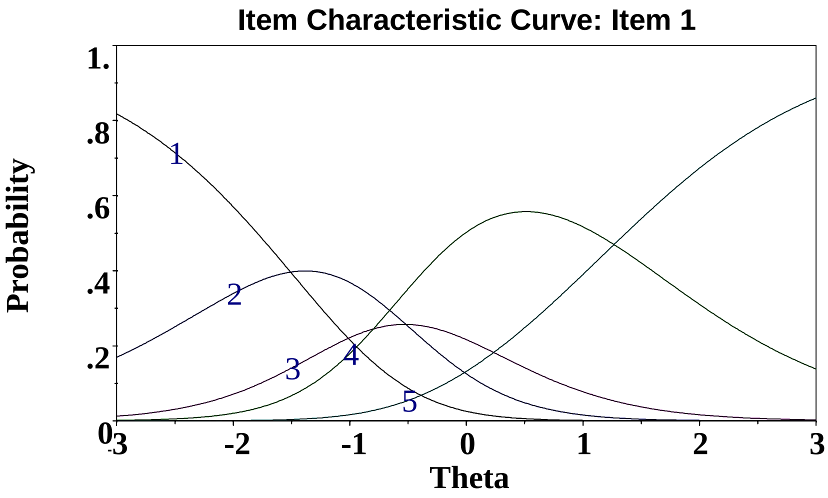


Figure 4.46

*Category Probability Plot for Item 2 with PCM: Sample 1, Leadership Scale*

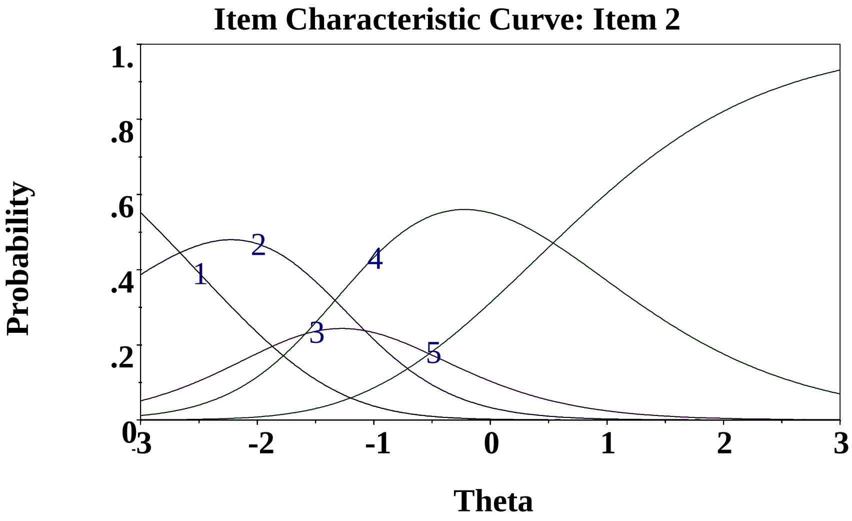


Figure 4.47

*Category Probability Plot for Item 3 with PCM: Sample 1, Leadership Scale*

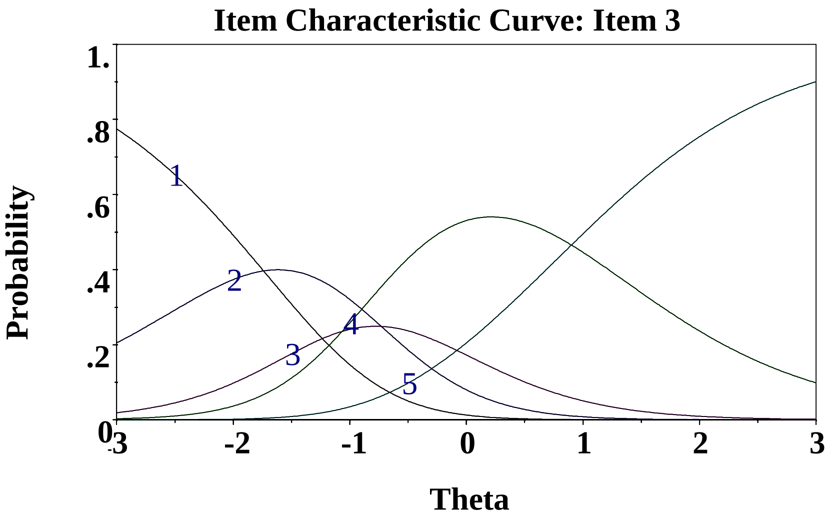


Figure 4.48

*Category Probability Plot for Item 4 with PCM: Sample 1, Leadership Scale*

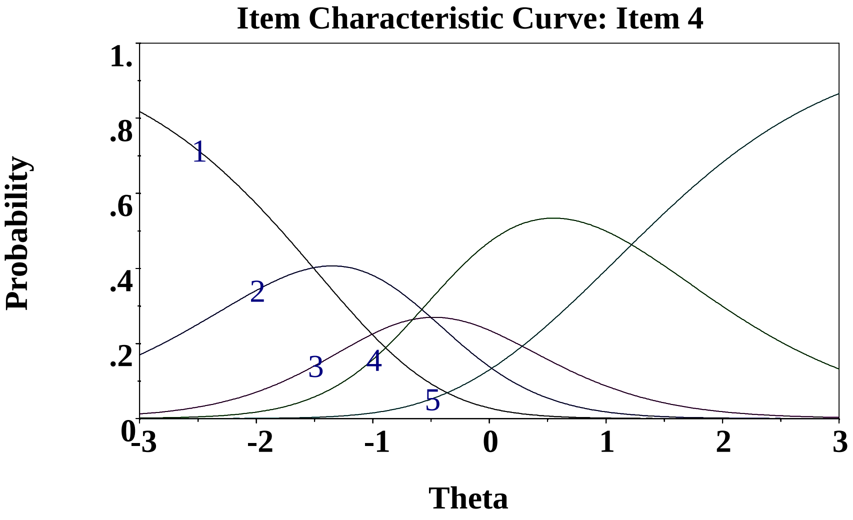


Figure 4.49

*Category Probability Plot for Item 5 with PCM: Sample 1, Leadership Scale*

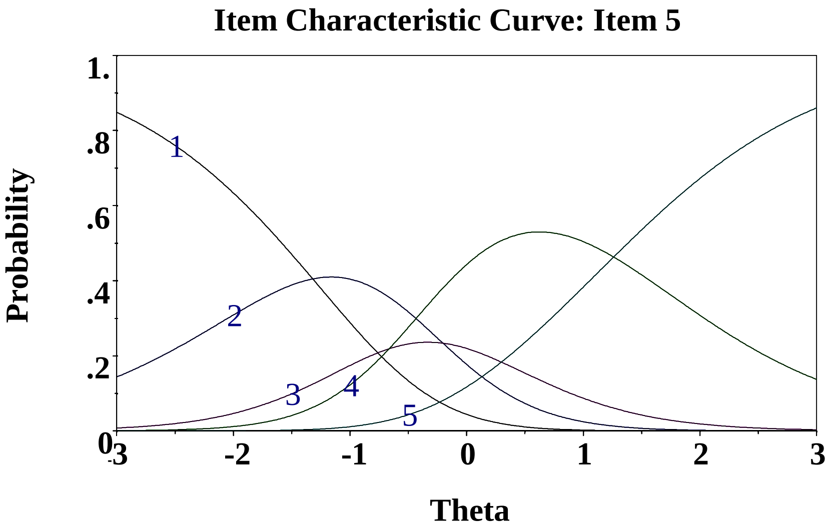


Figure 4.50

*Category Probability Plot for Item 6 with PCM: Sample 1, Leadership Scale*

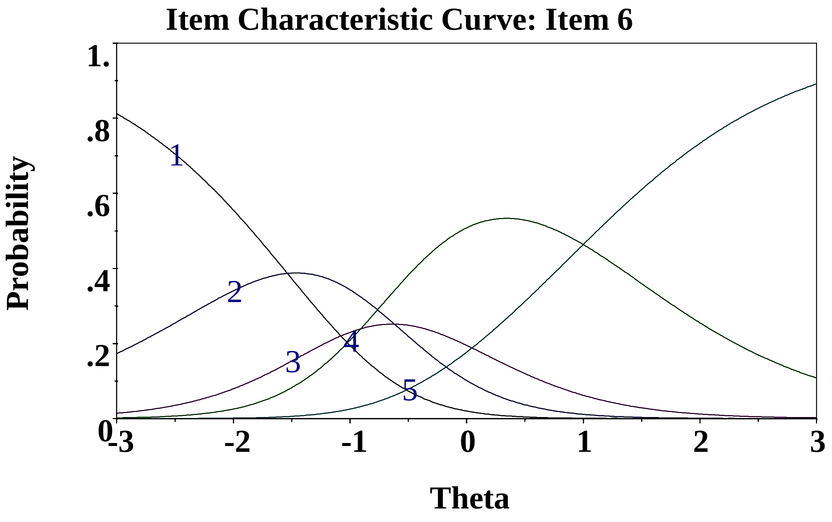


Figure 4.51

*Category Probability Plot for Item 7 with PCM: Sample 1, Leadership Scale*

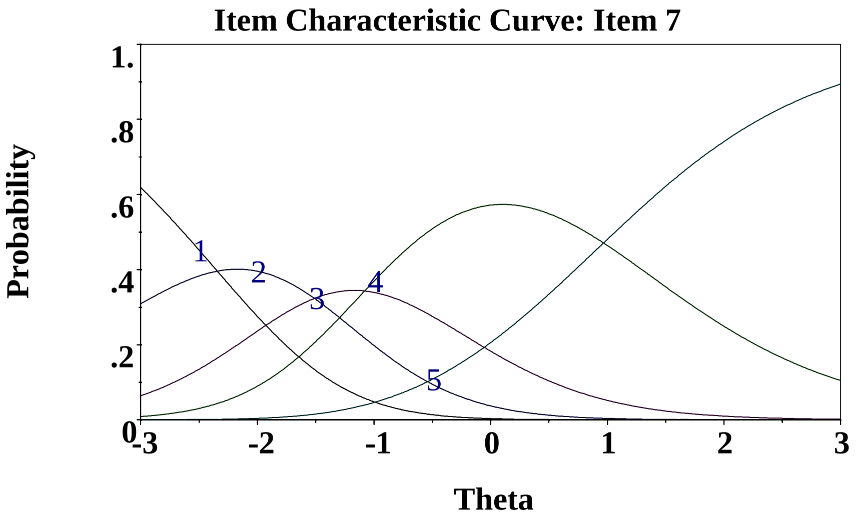


Figure 4.52

*Category Probability Plot for Item 8 with PCM: Sample 1, Leadership Scale*

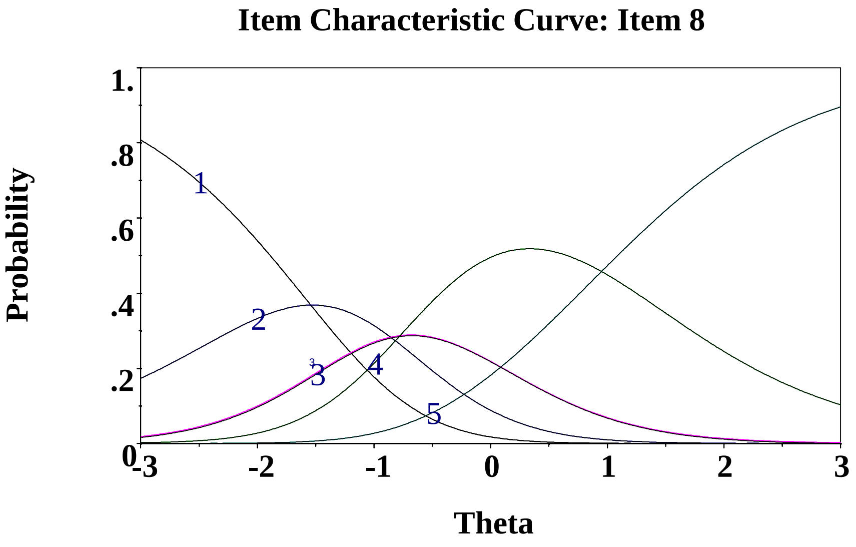


Figure 4.53

*Category Probability Plot for Item 9 with PCM: Sample 1, Leadership Scale*

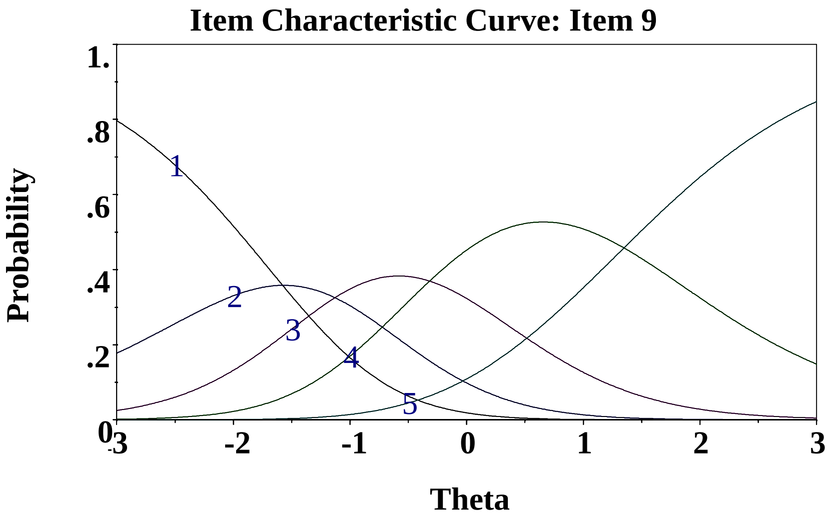


Figure 4.54

*Category Probability Plot for Item 10 with PCM: Sample 1, Leadership Scale*

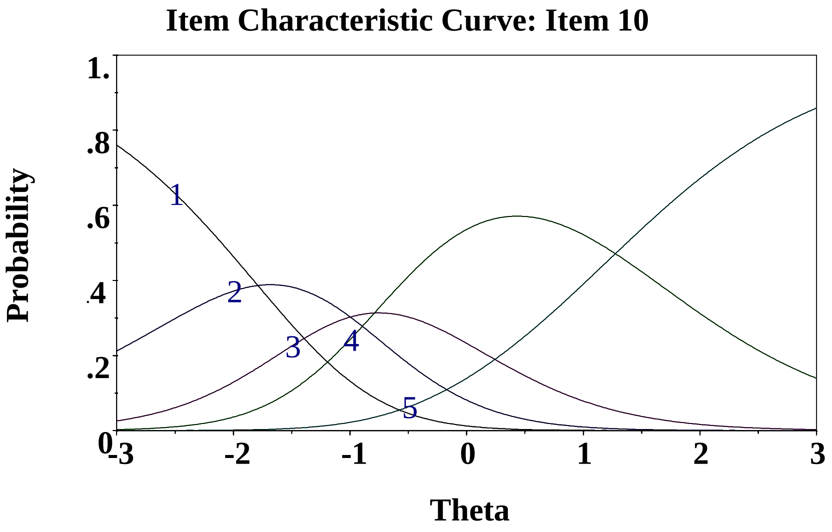


Figure 4.55

*Category Probability Plot for Item 11 with PCM: Sample 1, Leadership Scale*

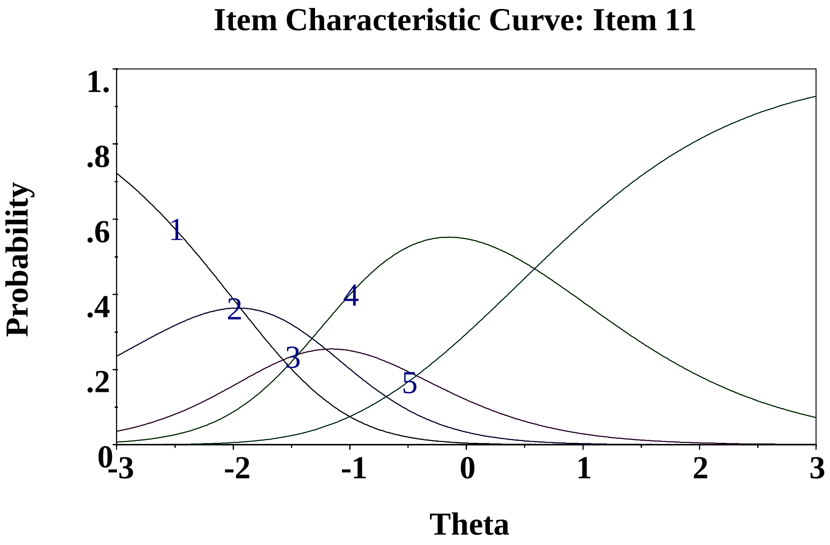


Figure 4.56

*Category Probability Plot for Item 12 with PCM: Sample 1, Leadership Scale*

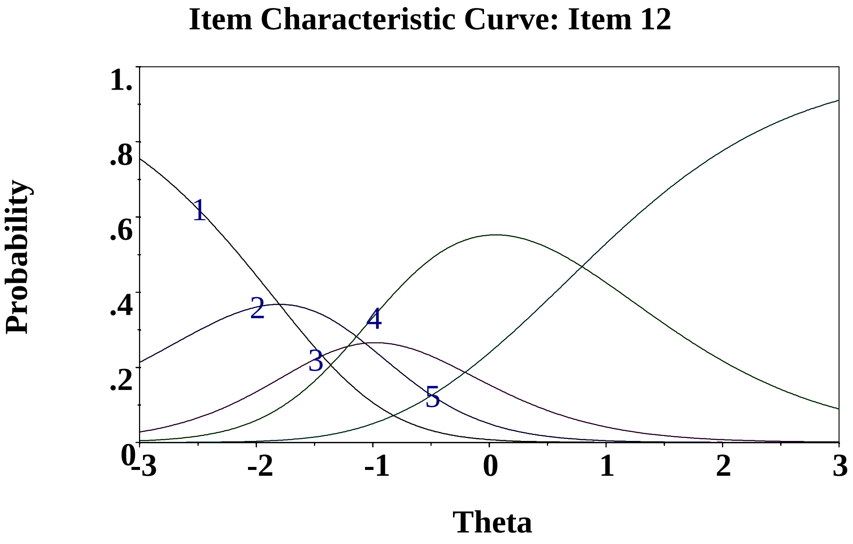


Figure 4.57

*Category Probability Plot for Item 13 with PCM: Sample 1, Leadership Scale*

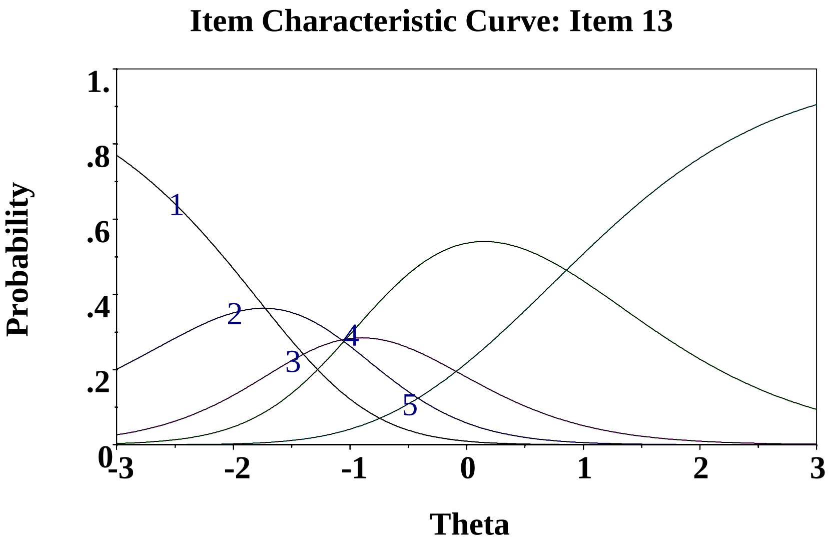


Figure 4.58

*Category Probability Plot for Item 14 with PCM: Sample 1, Leadership Scale*

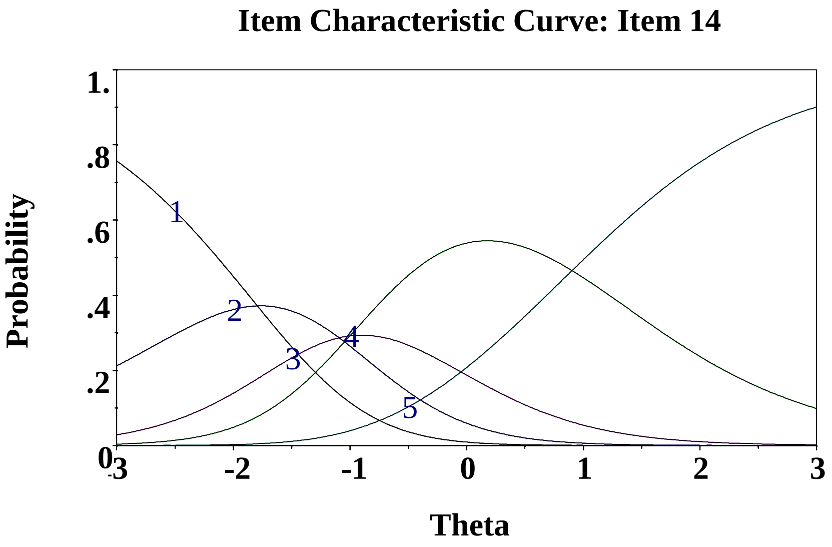


Figure 4.59

*Category Probability Plot for Item 15 with PCM: Sample 1, Leadership Scale*

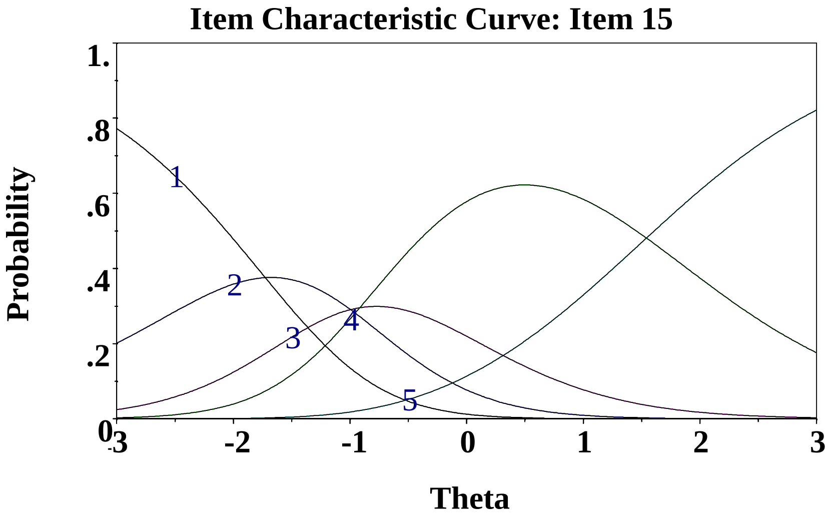


Figure 4.60

*Category Probability Plot for Item 16 with PCM: Sample 1, Leadership Scale*

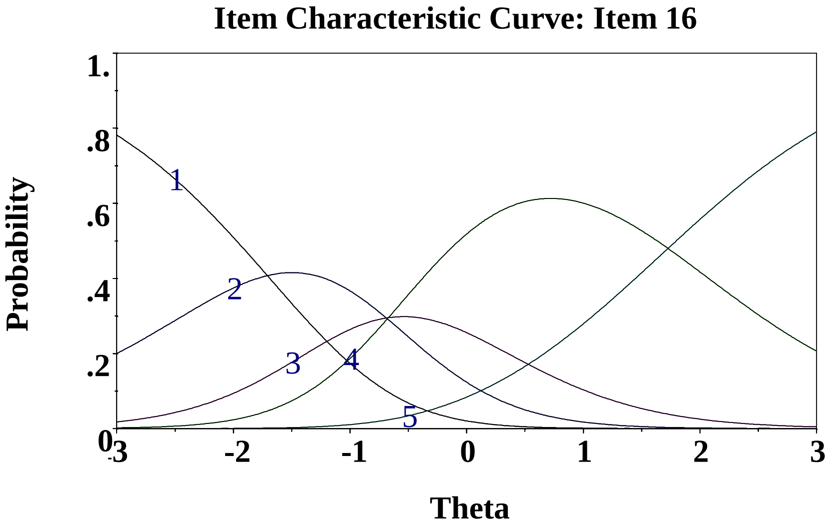


Figure 4.61

*Category Probability Plot for Item 17 with PCM: Sample 1, Leadership Scale*

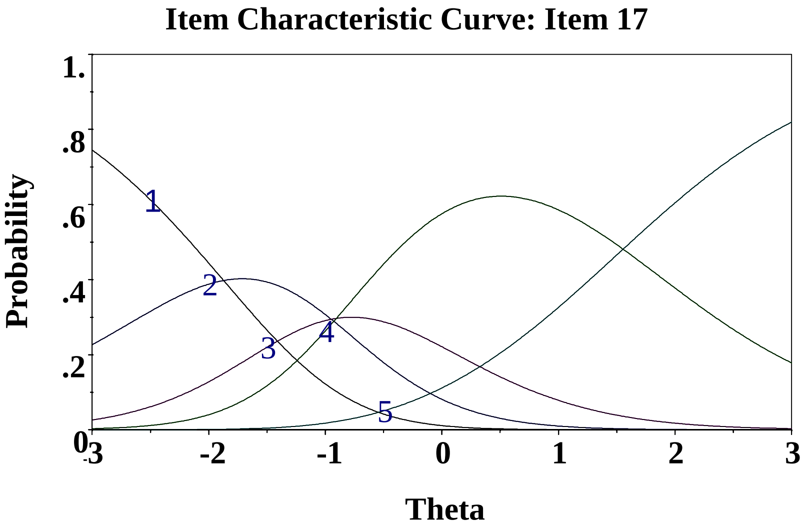


Figure 4.62

*Category Probability Plot for Item 18 with PCM: Sample 1, Leadership Scale*

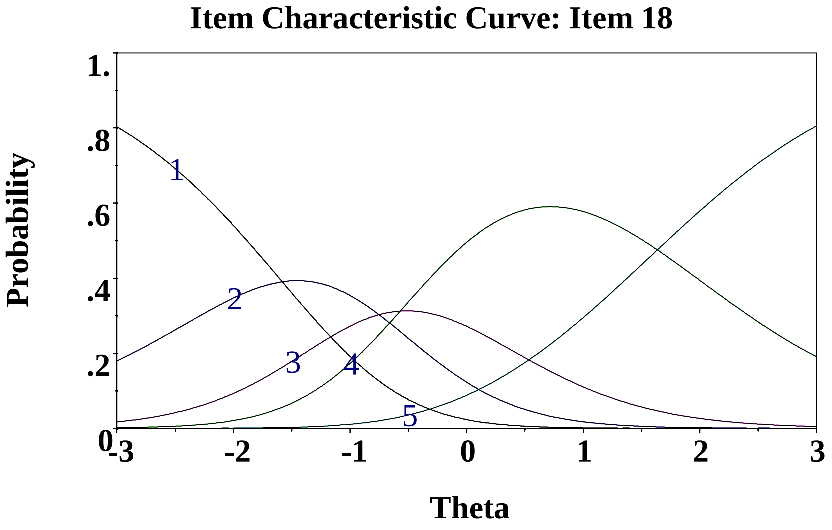


Figure 4.63

*Category Probability Plot for Item 19 with PCM: Sample 1, Leadership Scale*

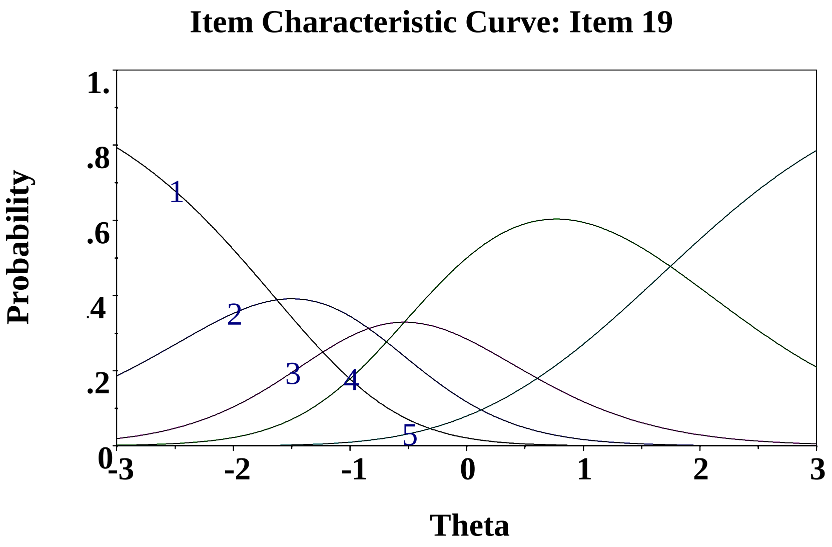


Figure 4.64

*Category Probability Plot for Item 20 with PCM: Sample 1, Leadership Scale*

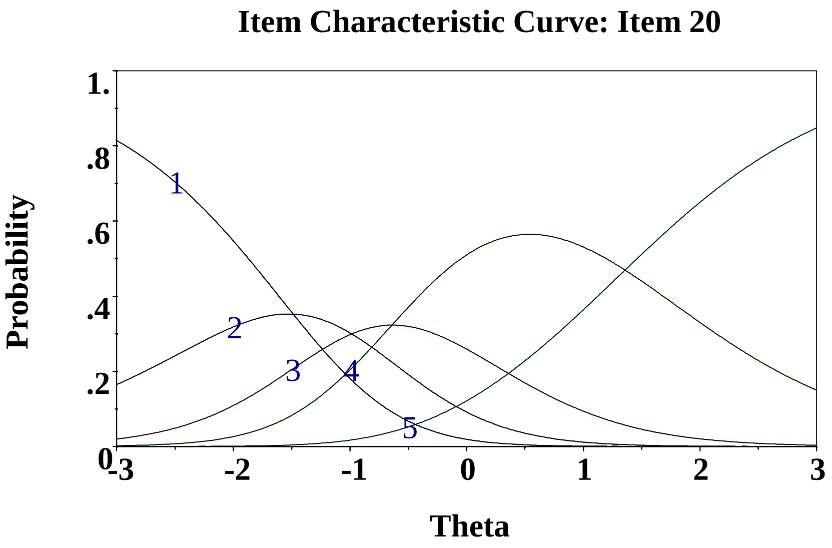


Figure 4.65

*Category Probability Plot for Item 21 with PCM: Sample 1, Leadership Scale*

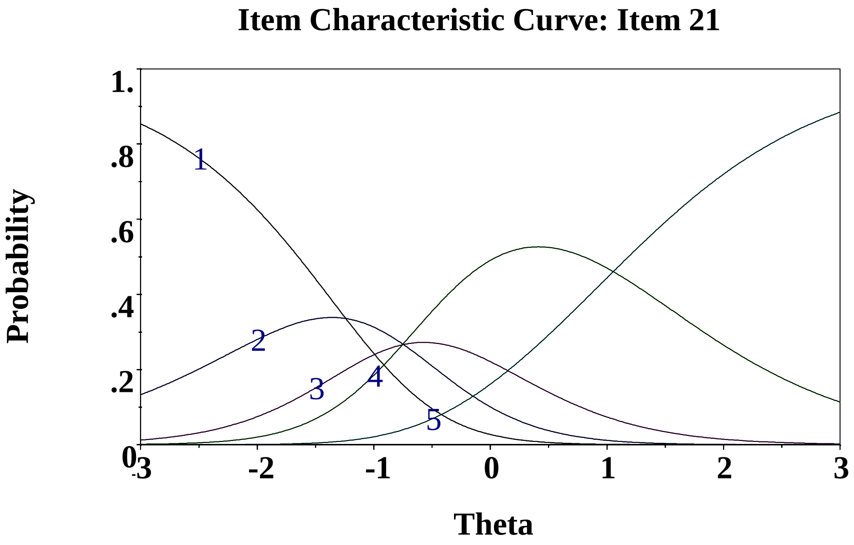


Figure 4.66

*Category Probability Plot for Item 1 with GPCM: Sample 1, Leadership Scale*

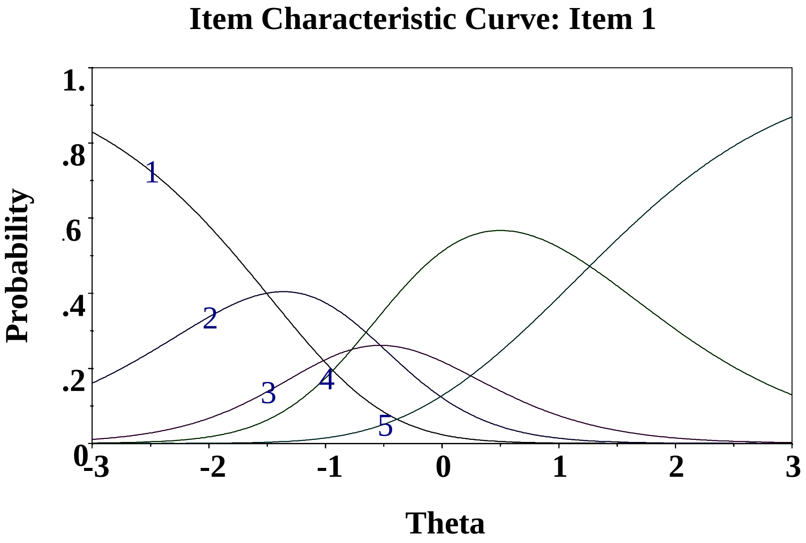


Figure 4.67

*Category Probability Plot for Item 2 with GPCM: Sample 1, Leadership Scale*

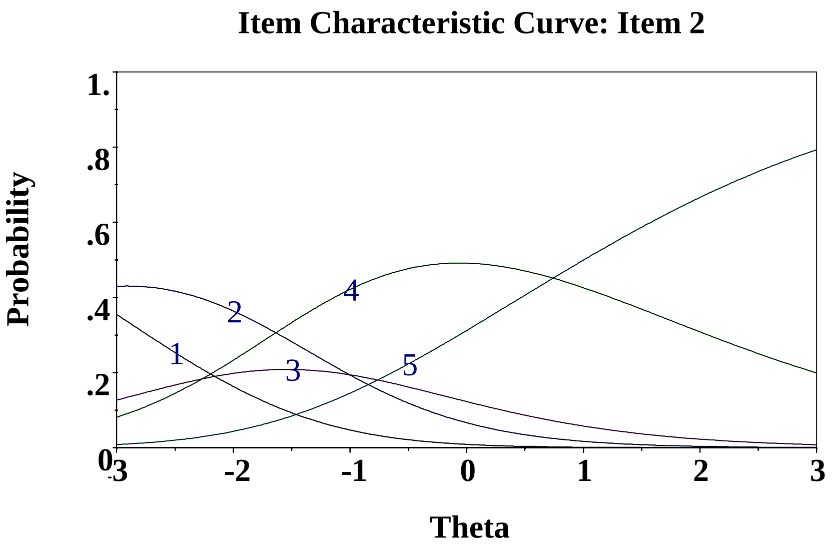


Figure 4.68

*Category Probability Plot for Item 3 with GPCM: Sample 1, Leadership Scale*

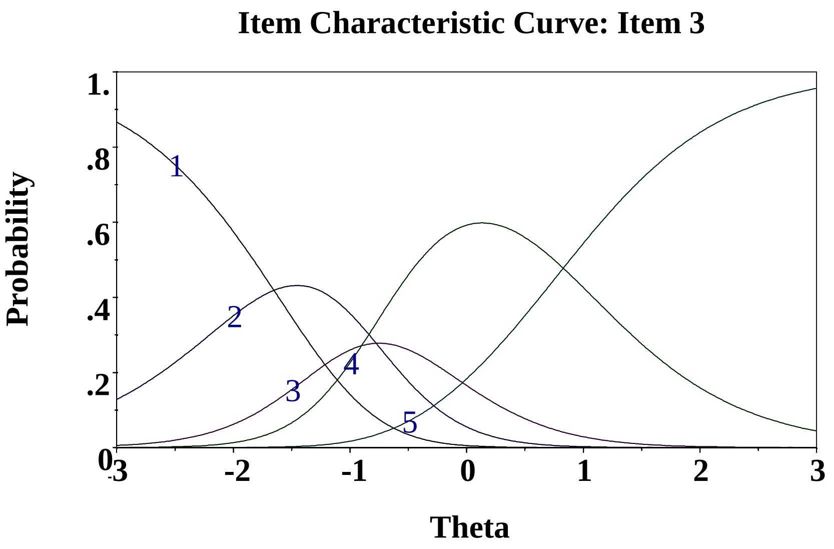


Figure 4.69

*Category Probability Plot for Item 4 with GPCM: Sample 1, Leadership Scale*

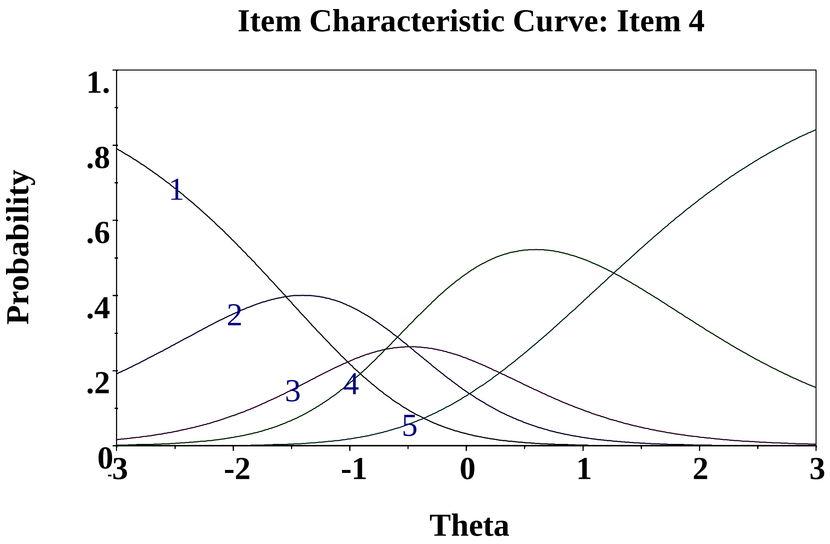


Figure 4.70

*Category Probability Plot for Item 5 with GPCM: Sample 1, Leadership Scale*

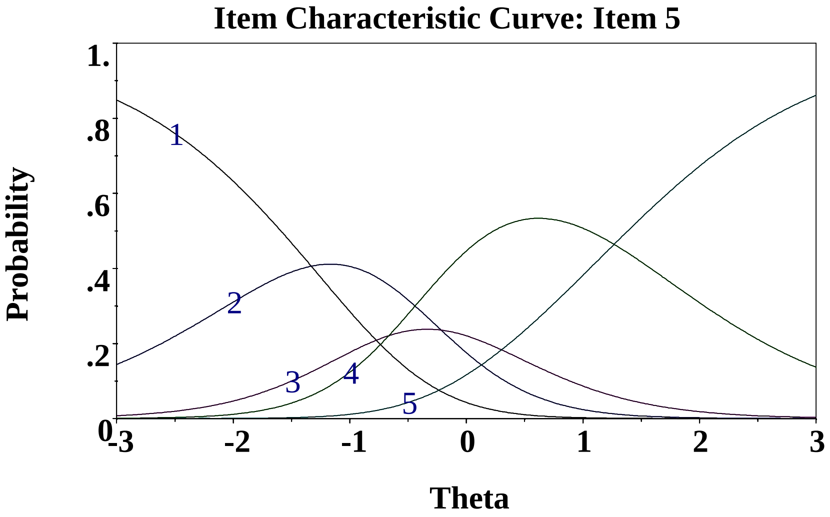


Figure 4.71

*Category Probability Plot for Item 6 with GPCM: Sample 1, Leadership Scale*

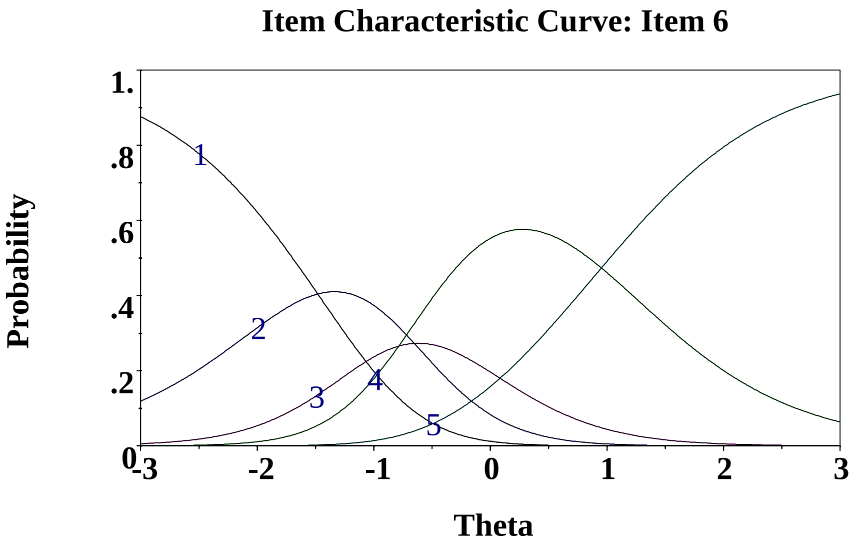


Figure 4.72

*Category Probability Plot for Item 7 with GPCM: Sample 1, Leadership Scale*

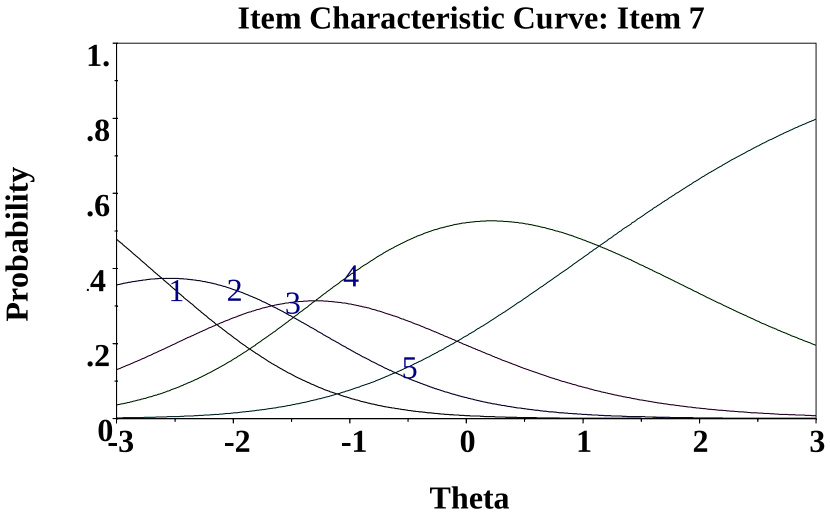


Figure 4.73

*Category Probability Plot for Item 8 with GPCM: Sample 1, Leadership Scale*

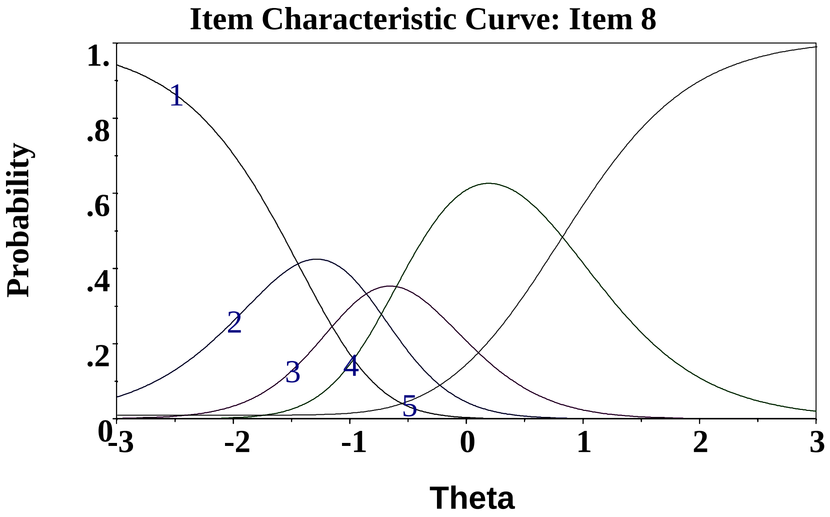


Figure 4.74

*Category Probability Plot for Item 9 with GPCM: Sample 1, Leadership Scale*

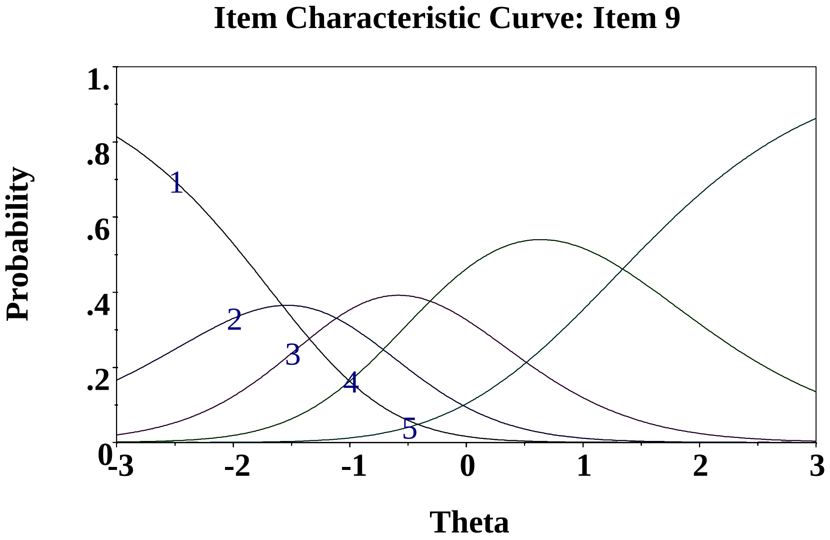


Figure 4.75

*Category Probability Plot for Item 10 with GPCM: Sample 1, Leadership Scale*

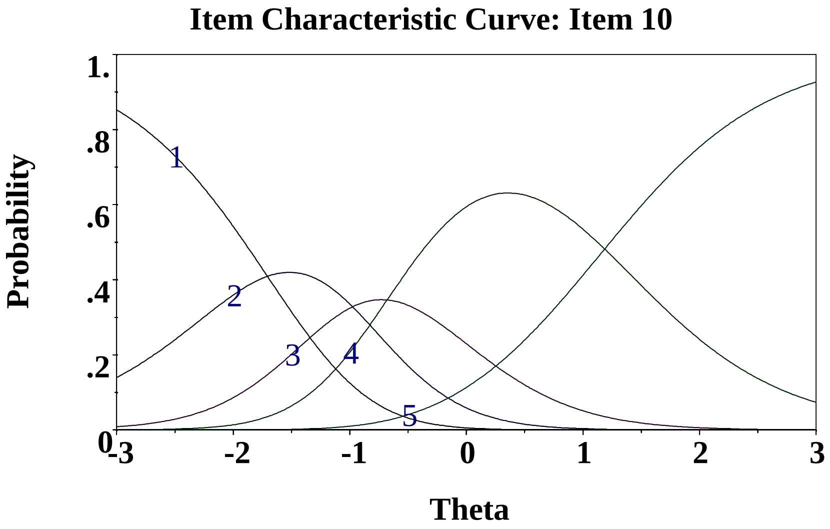


Figure 4.76

*Category Probability Plot for Item 11 with GPCM: Sample 1, Leadership Scale*

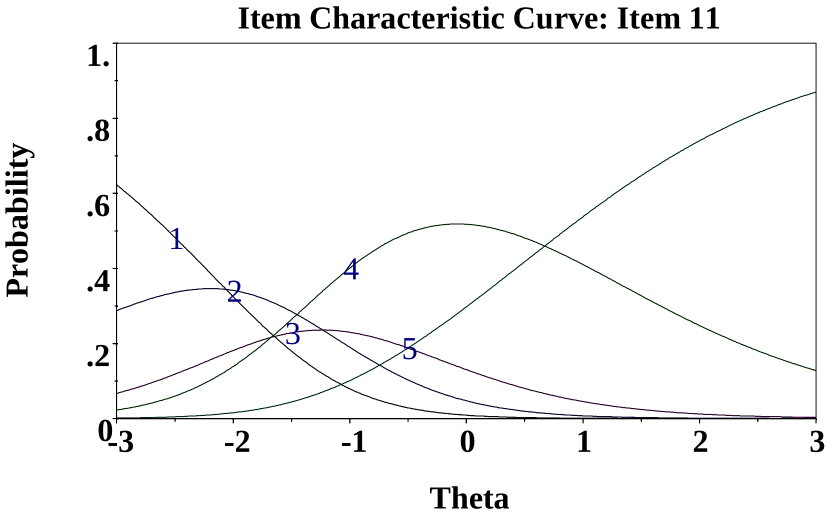


Figure 4.77

*Category Probability Plot for Item 12 with GPCM: Sample 1, Leadership Scale*

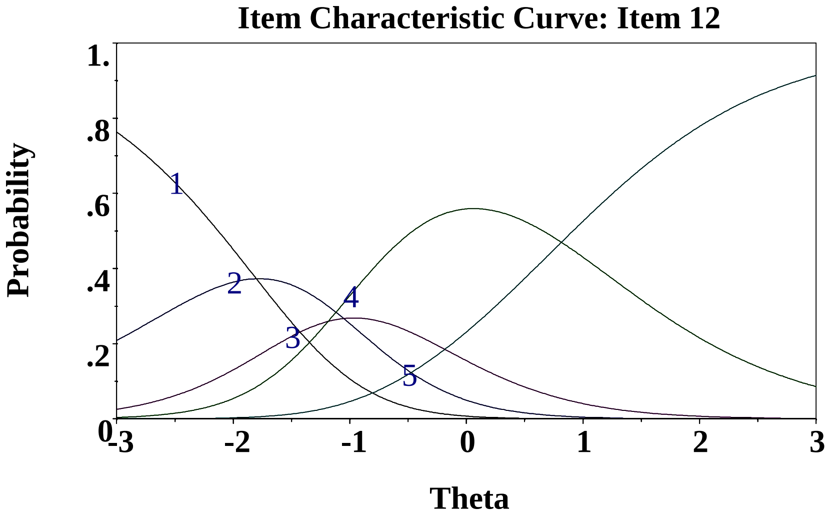


Figure 4.78

*Category Probability Plot for Item 13 with GPCM: Sample 1, Leadership Scale*

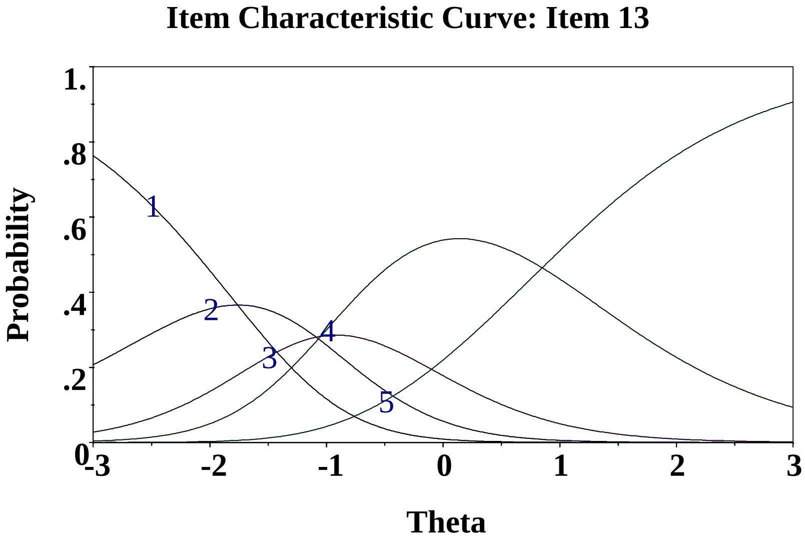


Figure 4.79

*Category Probability Plot for Item 14 with GPCM: Sample 1, Leadership Scale*

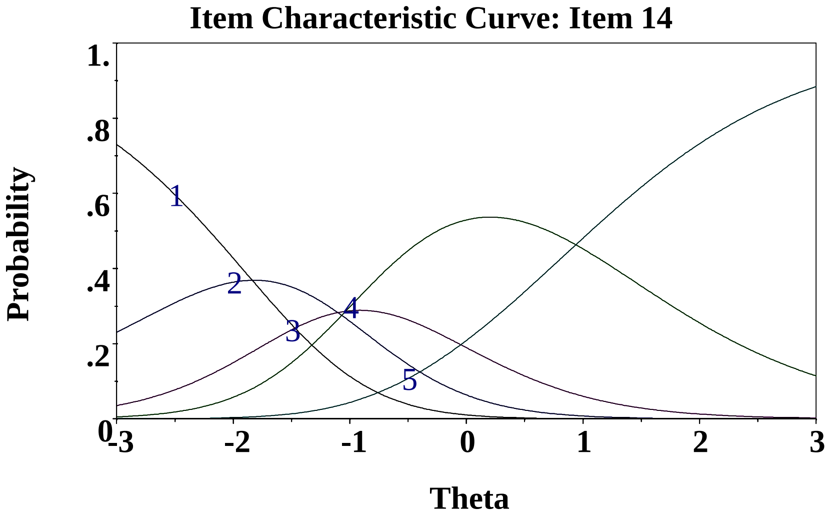


Figure 4.80

*Category Probability Plot for Item 15 with GPCM: Sample 1, Leadership Scale*

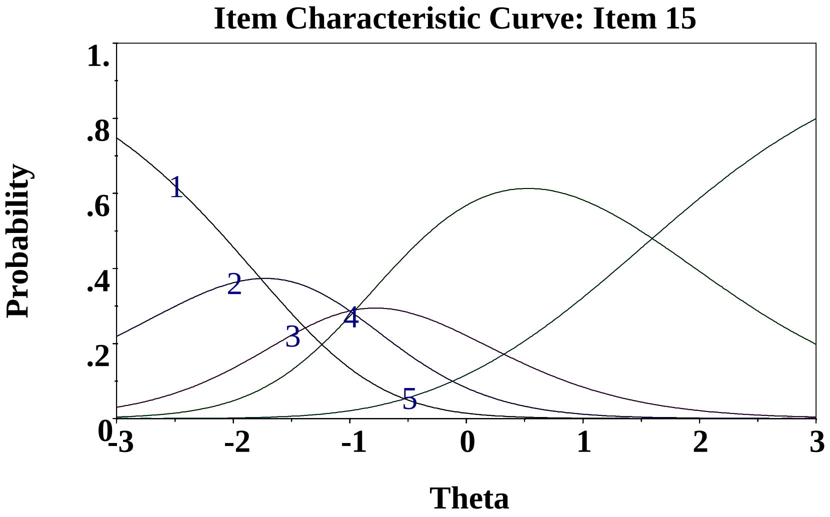


Figure 4.81

*Category Probability Plot for Item 16 with GPCM: Sample 1, Leadership Scale*

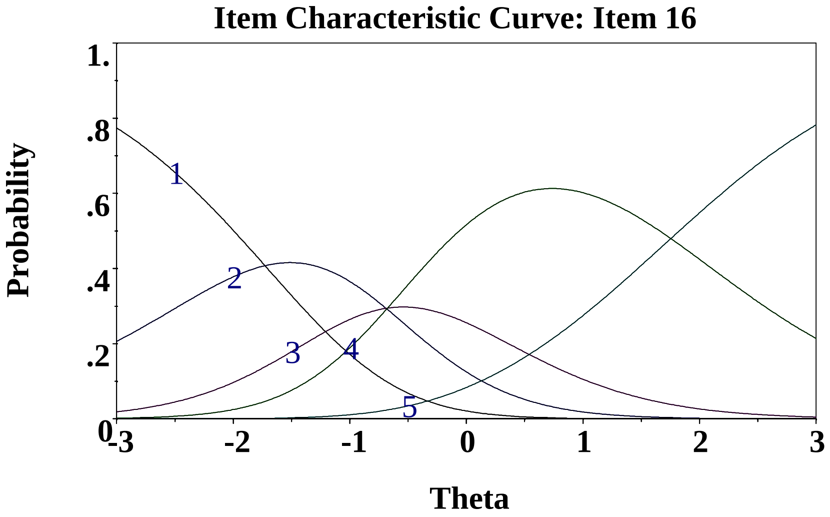


Figure 4.82

*Category Probability Plot for Item 17 with GPCM: Sample 1, Leadership Scale*

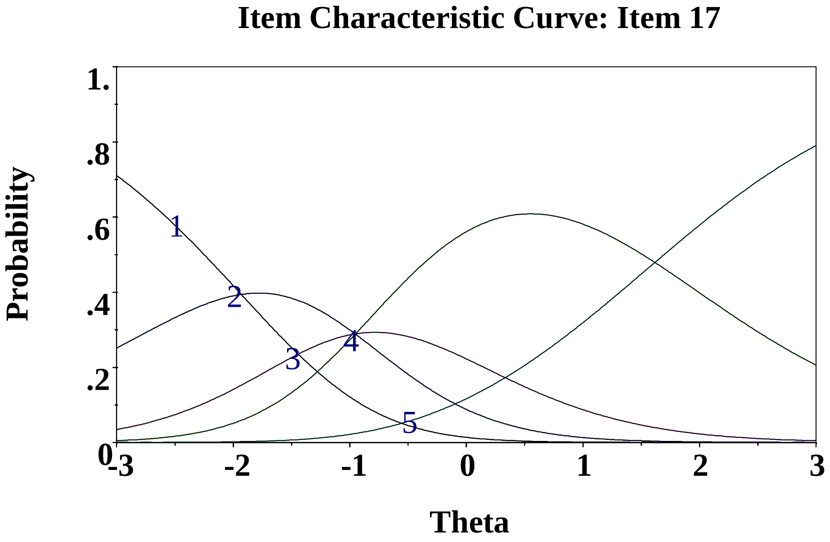


Figure 4.83

*Category Probability Plot for Item 18 with GPCM: Sample 1, Leadership Scale*

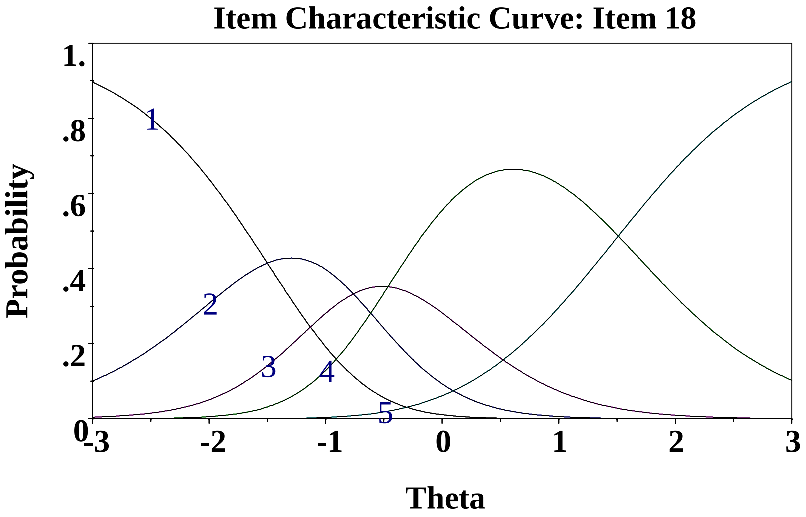


Figure 4.84

*Category Probability Plot for Item 19 with GPCM: Sample 1, Leadership Scale*

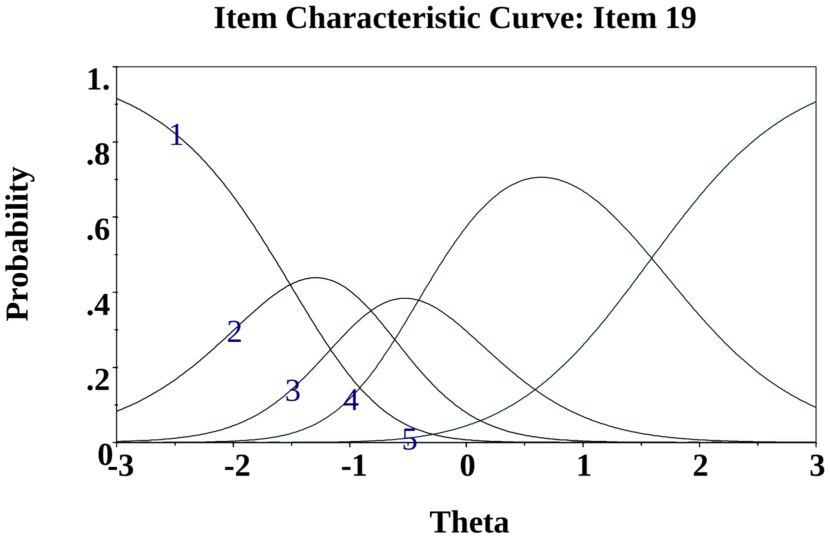


Figure 4.85

*Category Probability Plot for Item 20 with GPCM: Sample 1, Leadership Scale*

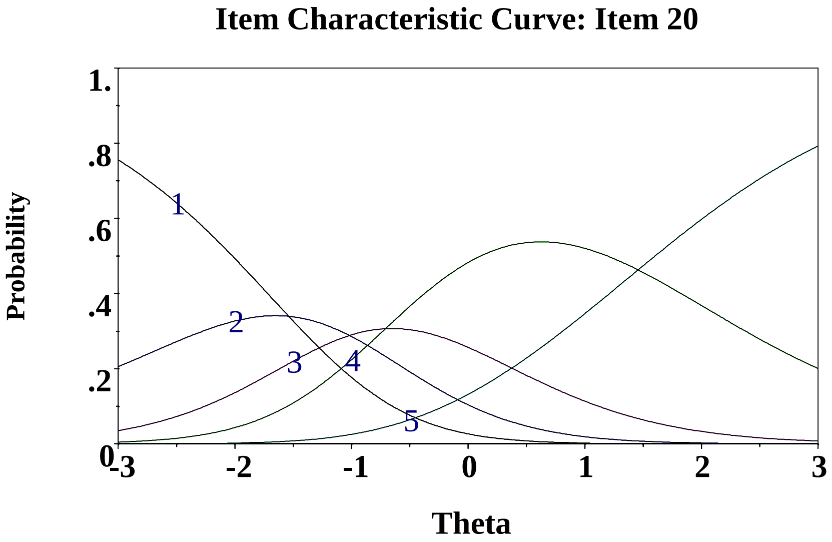


Figure 4.86

*Category Probability Plot for Item 21 with GPCM: Sample 1, Leadership Scale*

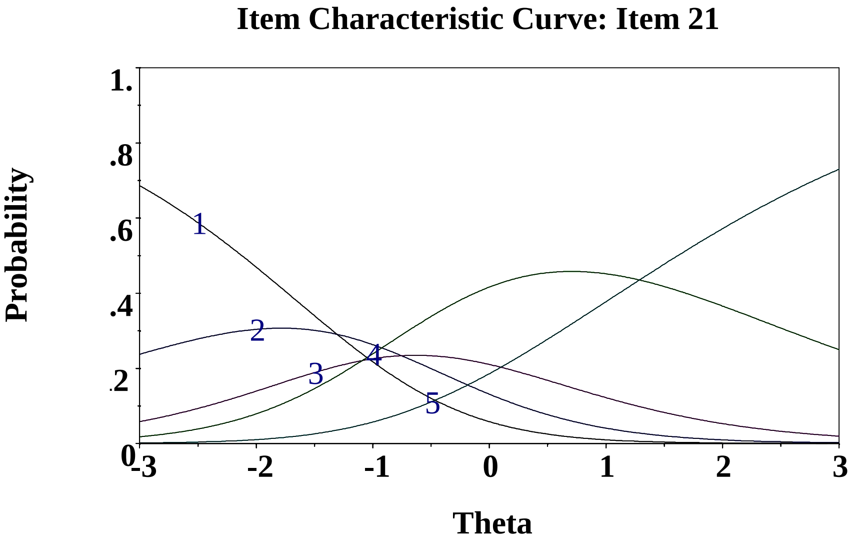


Figure 4.87

*Category Probability Plot for Item 1 with GGUM: Sample 1, Leadership Scale*

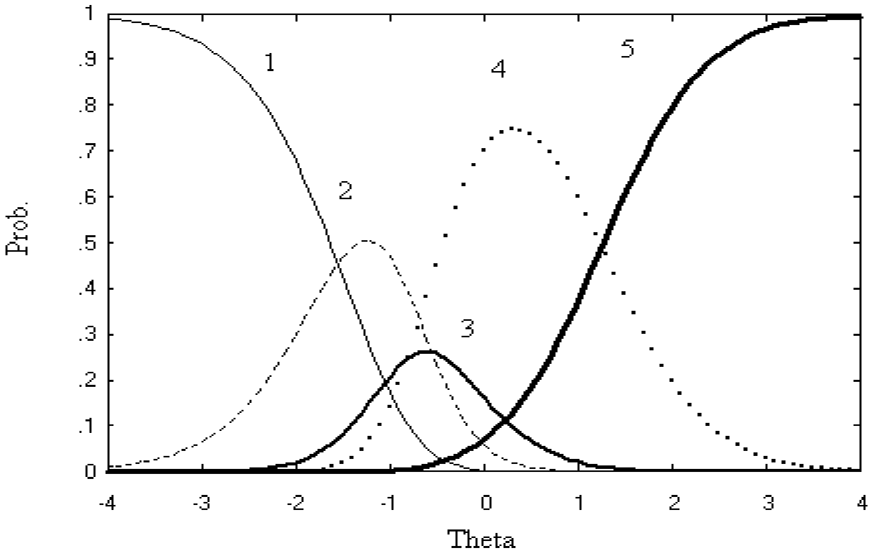


Figure 4.88

*Category Probability Plot for Item 2 with GGUM: Sample 1, Leadership Scale*

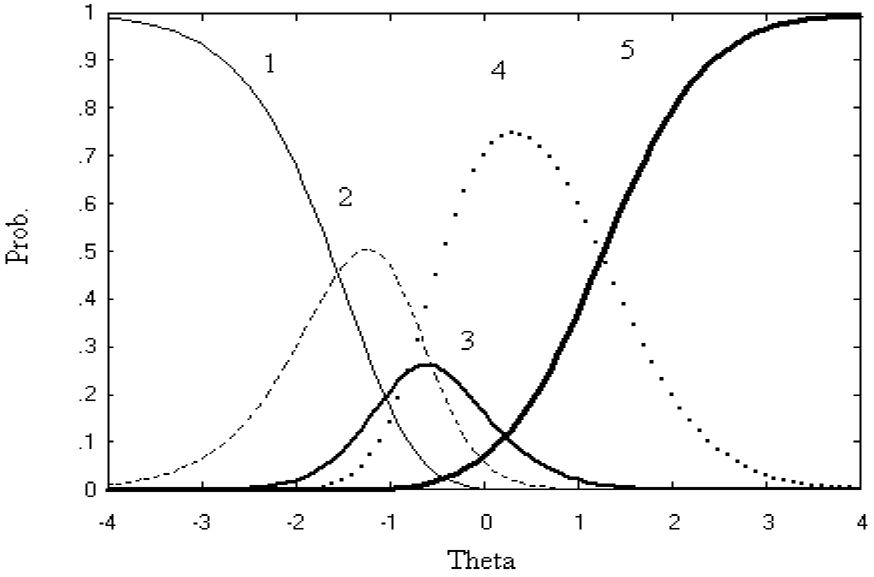


Figure 4.89

*Category Probability Plot for Item 3 with GGUM: Sample 1, Leadership Scale*

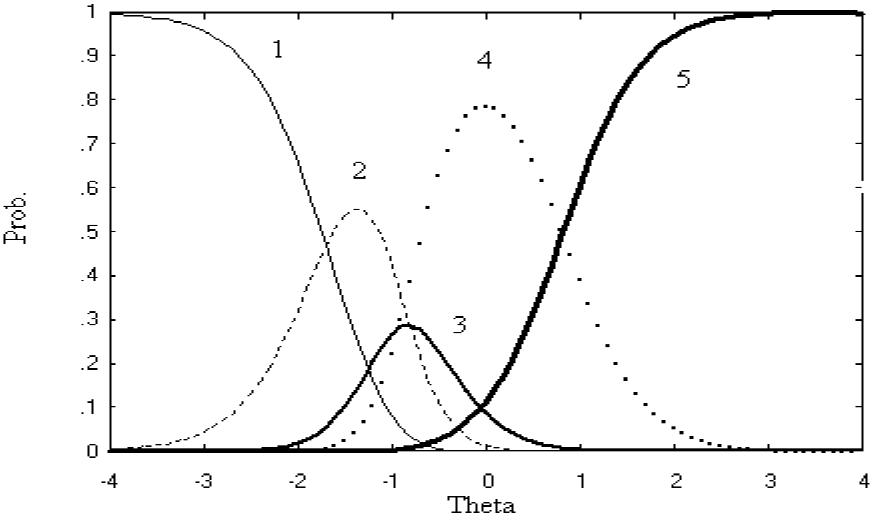


Figure 4.90

*Category Probability Plot for Item 4 with GGUM: Sample 1, Leadership Scale*

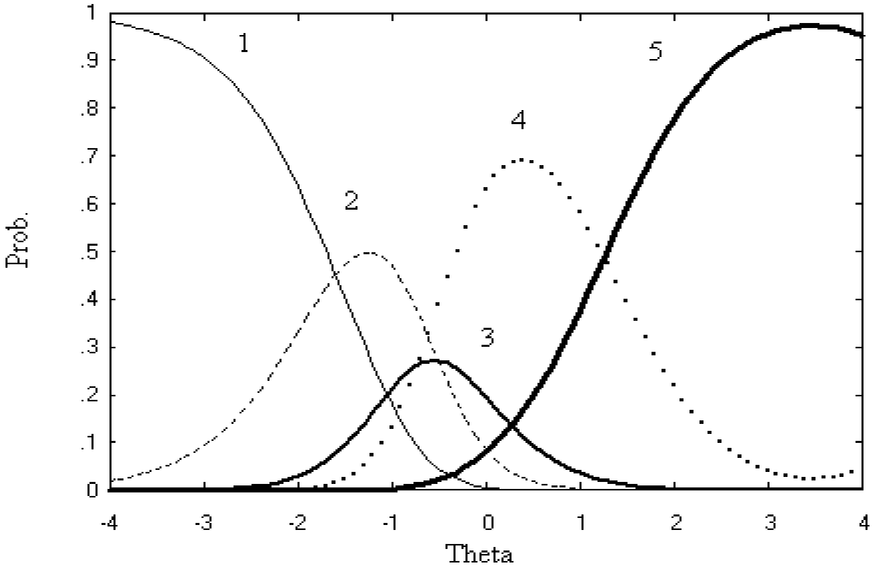


Figure 4.91

*Category Probability Plot for Item 5 with GGUM: Sample 1, Leadership Scale*

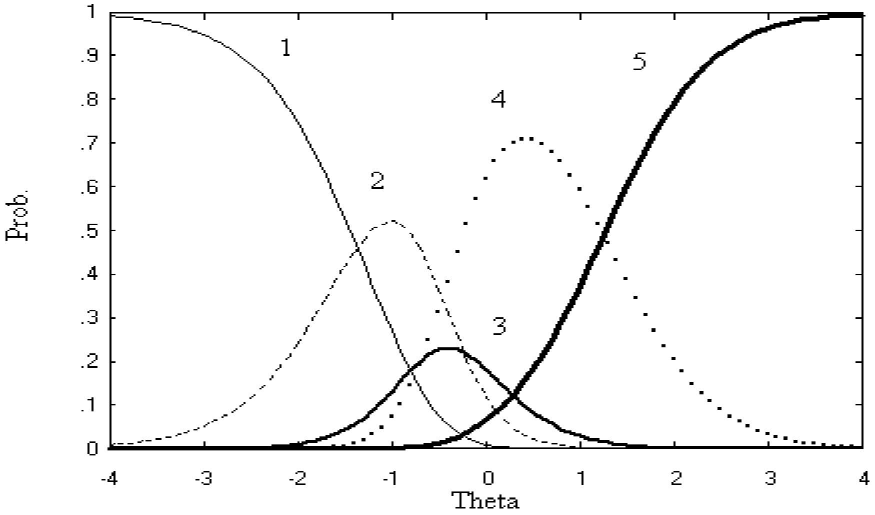


Figure 4.92

*Category Probability Plot for Item 6 with GGUM: Sample 1, Leadership Scale*

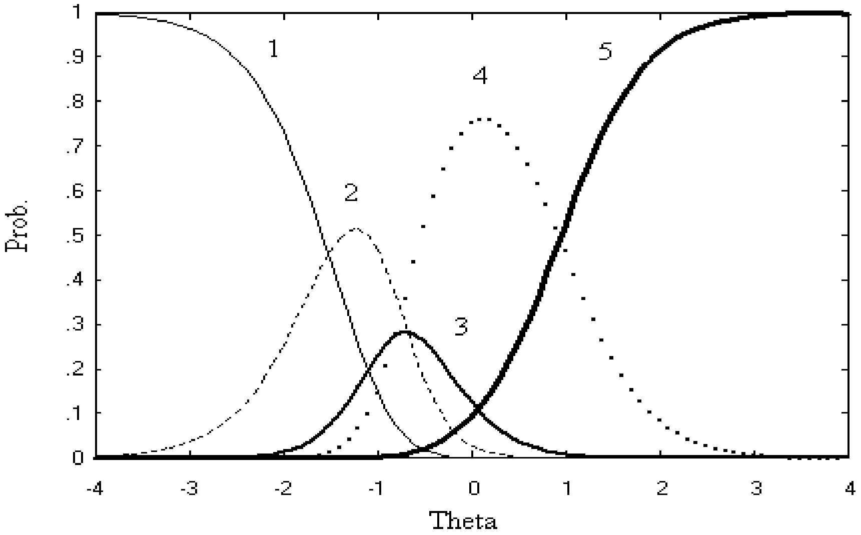


Figure 4.93

*Category Probability Plot for Item 7 with GGUM: Sample 1, Leadership Scale*

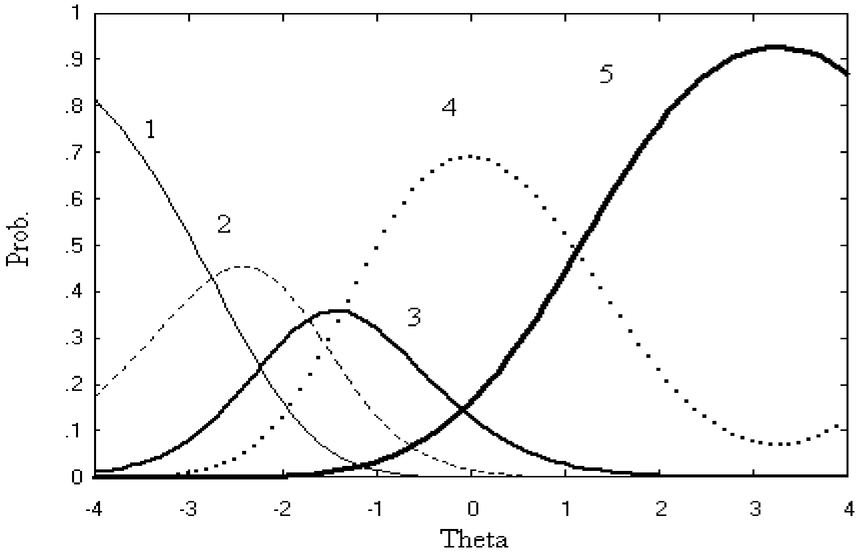


Figure 4.94

*Category Probability Plot for Item 8 with GGUM: Sample 1, Leadership Scale*

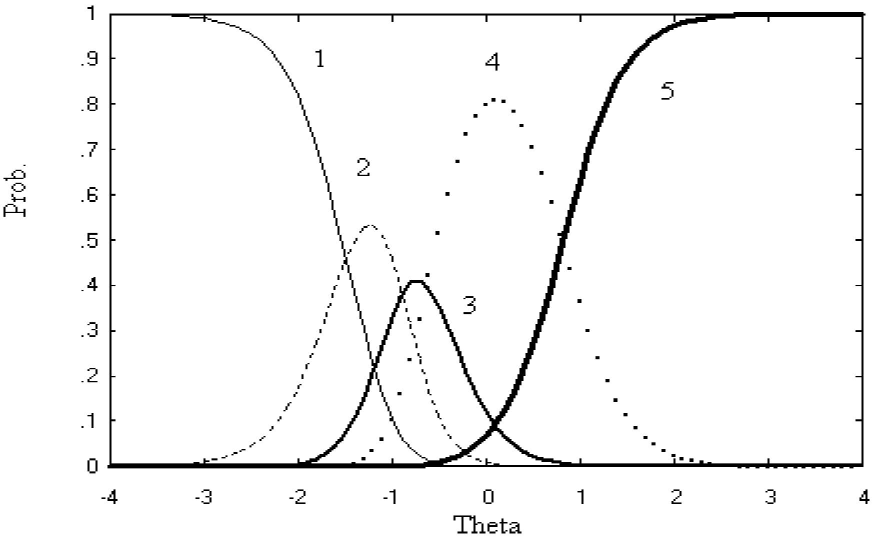


Figure 4.95

*Category Probability Plot for Item 9 with GGUM: Sample 1, Leadership Scale*

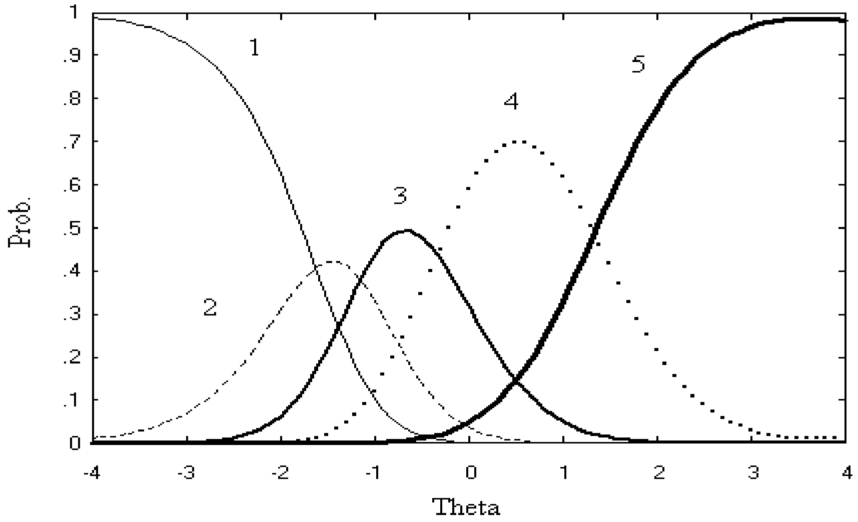


Figure 4.96

*Category Probability Plot for Item 10 with GGUM: Sample 1, Leadership Scale*

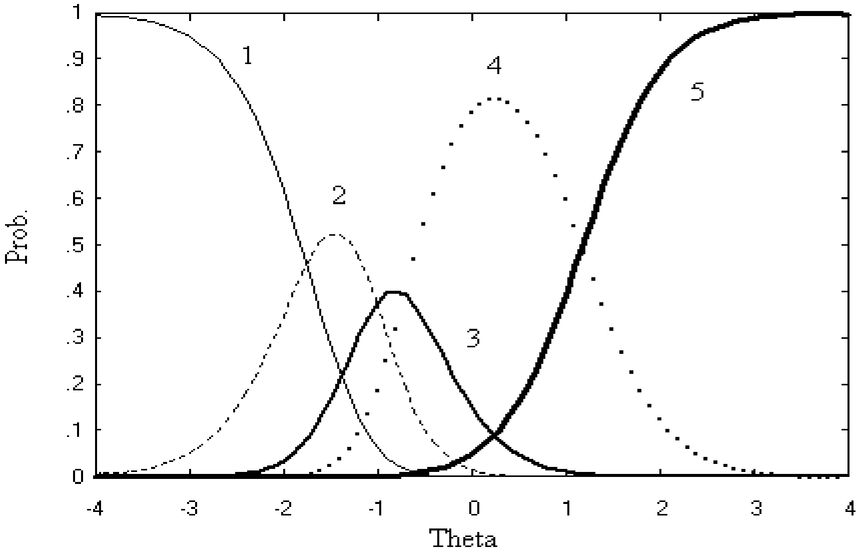


Figure 4.97

*Category Probability Plot for Item 11 with GGUM: Sample 1, Leadership Scale*

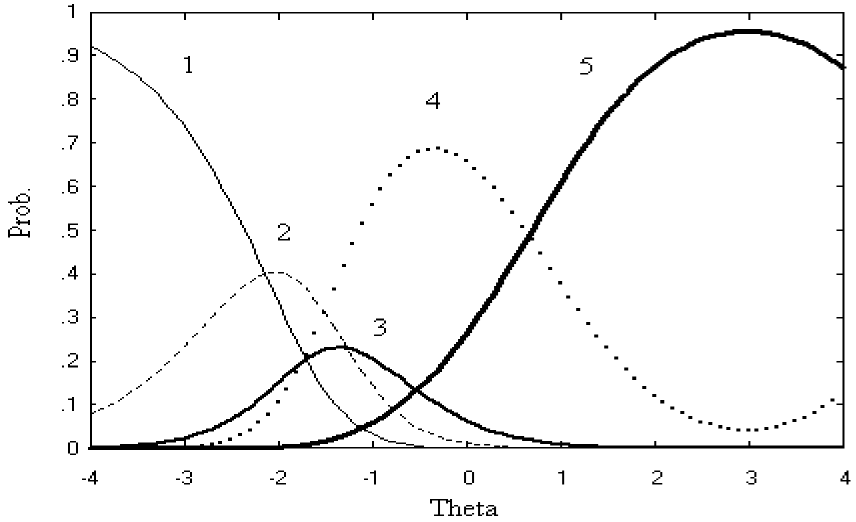


Figure 4.98

*Category Probability Plot for Item 12 with GGUM: Sample 1, Leadership Scale*

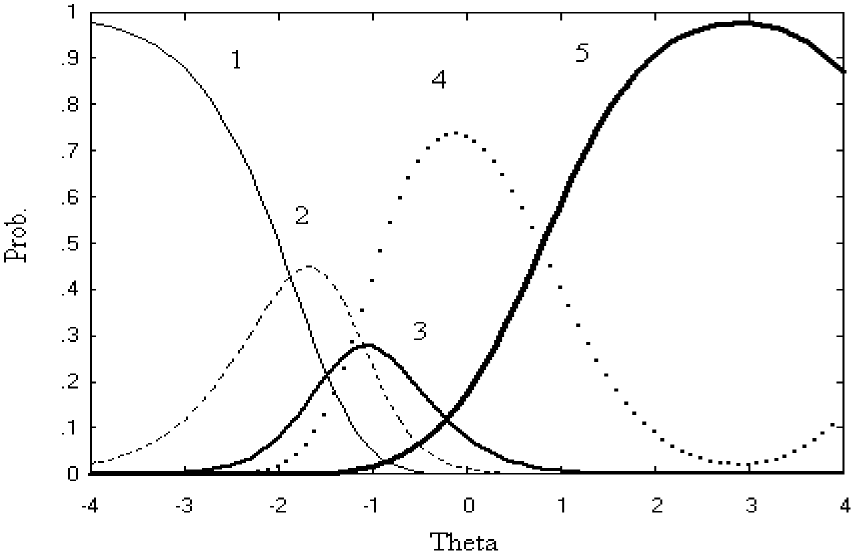


Figure 4.99

*Category Probability Plot for Item 13 with GGUM: Sample 1, Leadership Scale*

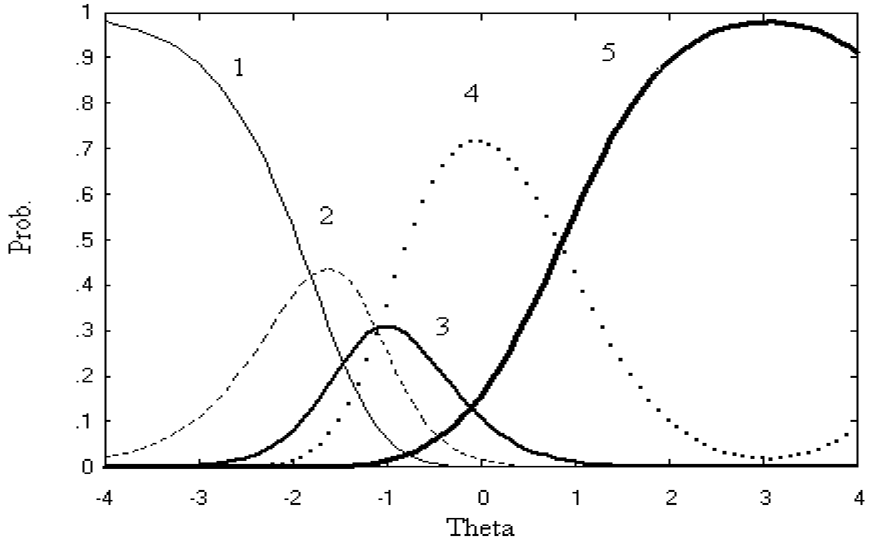


Figure 4.100

*Category Probability Plot for Item 14 with GGUM: Sample 1, Leadership Scale*

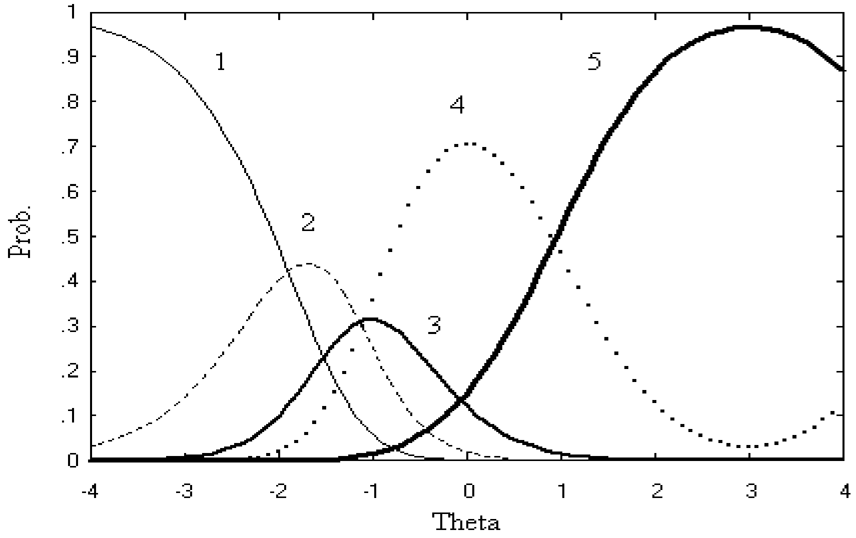


Figure 4.101

*Category Probability Plot for Item 15 with GGUM: Sample 1, Leadership Scale*

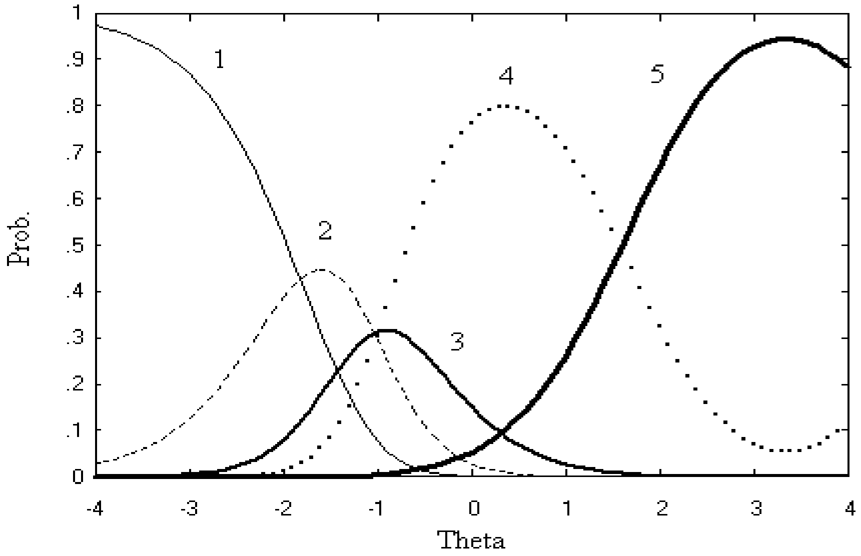


Figure 4.102

*Category Probability Plot for Item 16 with GGUM: Sample 1, Leadership Scale*

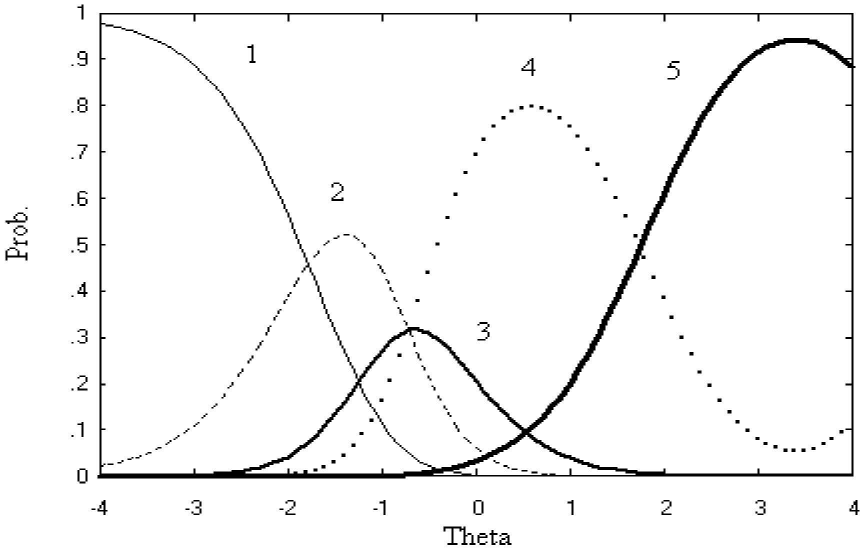


Figure 4.103

*Category Probability Plot for Item 17 with GGUM: Sample 1, Leadership Scale*

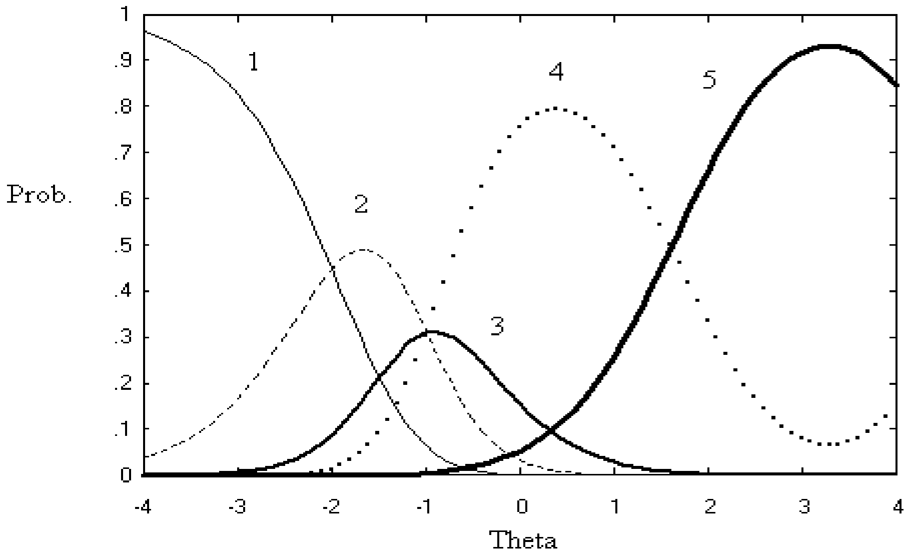


Figure 4.104

*Category Probability Plot for Item 18 with GGUM: Sample 1, Leadership Scale*

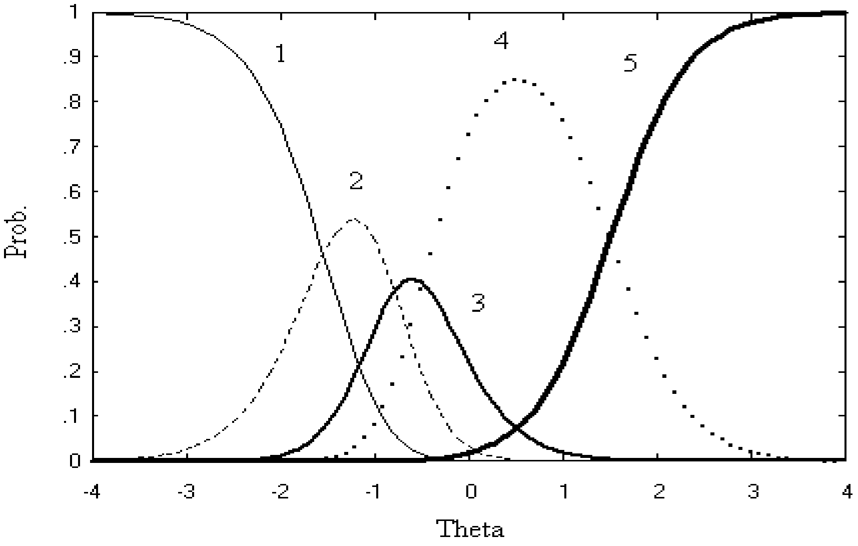


Figure 4.105

*Category Probability Plot for Item 19 with GGUM: Sample 1, Leadership Scale*

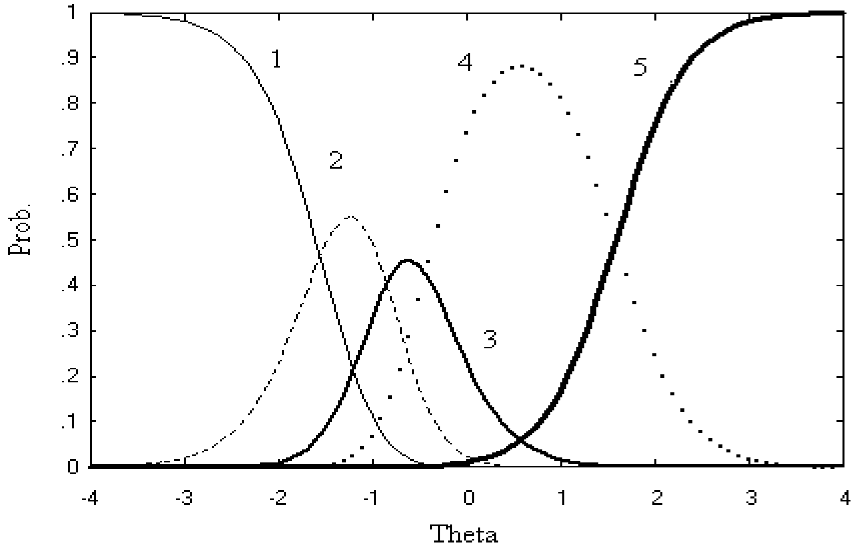


Figure 4.106

*Category Probability Plot for Item 20 with GGUM: Sample 1, Leadership Scale*

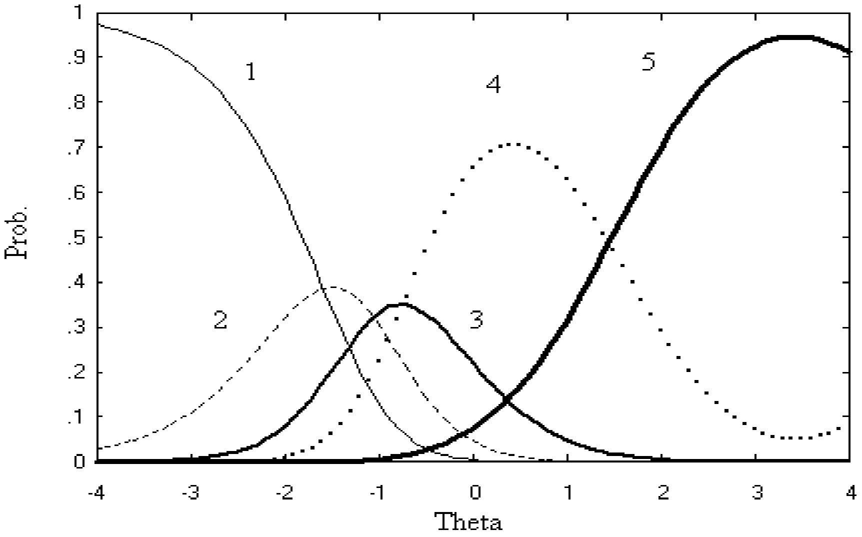
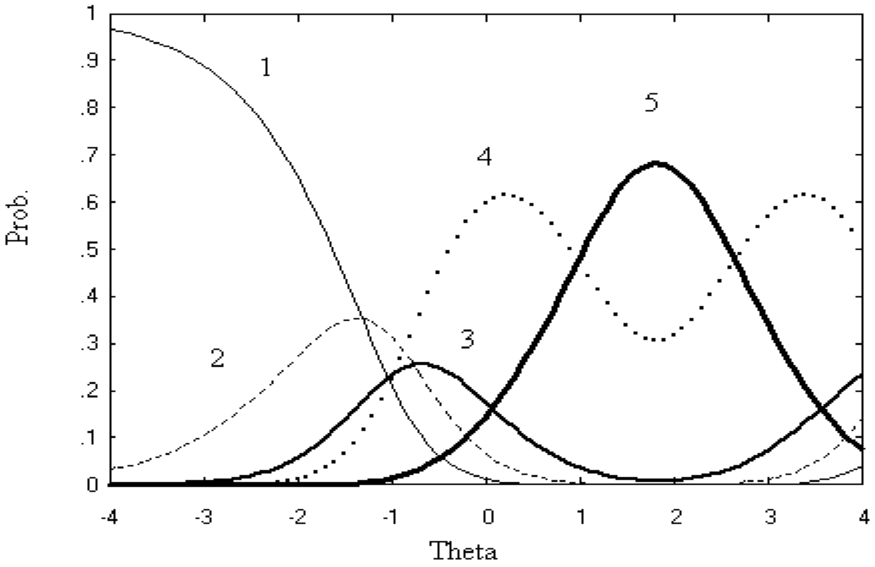


Figure 4.107

*Category Probability Plot for Item 21 with GGUM: Sample 1, Leadership Scale*



Just as in the Empowerment analyses, category 1 represented Strongly Disagree and category 5 represented Strongly Agree in all category probability plots. These plots are graphical representations of and reflect the item location, discrimination, and step/threshold item parameters. Because, on average, most respondents agreed with all of the items, because all 21 items were generally located in a very narrow region of the latent trait, and because the majority of the items discriminated similarly, the category probability functions across most of the 21 items looked very similar within and between analyses by IRT model.

Of note are the similarities between the cumulative and unfolding models. Based on examination of the category probability plots, there appears to be little difference between the two type of IRT models, with the exception of item 21. That item was least discriminating across the GPCM and GGUM models, though the rank order of the item's location was very different across the two model types. The GGUM estimated item 21 to be least extreme relative to the other items, and the cumulative models estimated this item to be moderate. Across all analyses, the category probability plot for item 21 is the only one that exhibited distinct unfolding properties. Specifically, the observed Strongly Agree response function is unidmodal and the observed other response functions are or approach bimodal. Interpretation of the Neither Agree Nor Disagree observed response function (i.e., category 3) is necessarily imprecise. This observed response function is the result of the two subjective response functions: Neither Agree Nor Disagree from above and Neither Agree Nor Disagree from below. Theoretically and practically, it is difficult to understand what an ambivalent response from above or from below mean. For item 21 in sample 1, the Neither Agree Nor Disagree operates similar to the Disagree response function. Items 7, and 11 through 17, also exhibit some unfolding properties, though only for the Agree response category, and slightly for the Strongly Agree category.

To address the third research question, the probability functions were generally very similar across the two types of IRT models, indicating little difference between cumulative and unfolding IRT models. Some items, did however, exhibit unfolding properties, especially for the Strongly Agree and Agree response options, generally, for items 7, 11 through 15 and 21. These properties were also evidenced by the slight non-monotonicity of the ICCs generated from the GGUM analyses for the two response options.

The second part of the fourth research question investigated in this study had to do with the fit of each model to the data. This was examined by calculating both absolute and relative fit statistics. Fit statistics are presented and discussed below.

Item and scale level chi-square distributed fit statistics ($G^2$) were calculated across the three IRT models. Results from the PCM analyses are reported in Table 4.32 for each of the 10 samples. Table 4.33 presents the fit statistics from the GPCM analyses, and Table 4.34 displays the fit statistics from the GGUM analyses. Ten fit groups were specified in all analyses. The *p*-values associated with an asterisk (*) indicate a statistically good fitting item at the .01 level. In the PCM analyses, item 1 showed significant fit in two of the ten samples, and item 4 exhibited good fit in a single sample. Across all analyses, the PCM never fit the survey statistically well as a whole. Based on the results in Table 4.32, the PCM is an ill-fitting model to this Leadership data. At the survey level, the fit of the GPCM model to the Leadership was not improved. Item level fit was marginally improved with the GPCM, compared to the PCM, with item 8 showing good fit in 5 of 10 analyses. Overall, results in Table 4.34 indicate that the GPCM is also an ill-fitting model for these data.

Table 4.32

*Item and Scale Level Chi-Square Fit Statistics for Each Leadership Sample: PCM*

| | Sample 1 | | | Sample 2 | | | Sample 3 | | | Sample 4 | | | Sample 5 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Item | $\chi^2$ | *df* | *p* | $\chi^2$ | *df* | *p* | $\chi^2$ | *df* | *p* | $\chi^2$ | *df* | *p* | $\chi^2$ | *df* | *p* |
| 1 | 61.924 | 23 | .00 | 40.492 | 21 | .01 | 48.923 | 22 | .00 | 53.716 | 20 | .00 | 57.801 | 21 | .00 |
| 2 | 239.075 | 18 | .00 | 244.923 | 18 | .00 | 159.432 | 18 | .00 | 179.428 | 18 | .00 | 266.817 | 18 | .00 |
| 3 | 110.603 | 21 | .00 | 133.353 | 22 | .00 | 99.433 | 21 | .00 | 85.667 | 19 | .00 | 82.972 | 19 | .00 |
| 4 | 48.462 | 23 | .00 | 31.768 | 22 | .08* | 60.313 | 21 | .00 | 53.050 | 21 | .00 | 46.697 | 21 | .00 |
| 5 | 67.569 | 23 | .00 | 97.385 | 25 | .00 | 47.708 | 21 | .00 | 42.277 | 21 | .00 | 90.786 | 21 | .00 |
| 6 | 87.549 | 21 | .00 | 67.194 | 22 | .00 | 65.954 | 22 | .00 | 53.168 | 19 | .00 | 103.596 | 21 | .00 |
| 7 | 83.167 | 20 | .00 | 85.550 | 20 | .00 | 76.615 | 19 | .00 | 103.931 | 18 | .00 | 149.802 | 18 | .00 |
| 8 | 204.553 | 22 | .00 | 178.663 | 22 | .00 | 184.111 | 22 | .00 | 192.598 | 19 | .00 | 197.778 | 21 | .00 |
| 9 | 77.925 | 23 | .00 | 109.119 | 24 | .00 | 62.795 | 21 | .00 | 90.912 | 22 | .00 | 70.221 | 22 | .00 |
| 10 | 96.768 | 23 | .00 | 87.111 | 22 | .00 | 74.597 | 21 | .00 | 94.522 | 20 | .00 | 64.769 | 21 | .00 |
| 11 | 227.497 | 19 | .00 | 163.646 | 19 | .00 | 139.076 | 18 | .00 | 150.577 | 18 | .00 | 175.591 | 19 | .00 |
| 12 | 94.444 | 20 | .00 | 54.438 | 20 | .00 | 90.227 | 18 | .00 | 89.302 | 19 | .00 | 86.350 | 18 | .00 |
| 13 | 76.184 | 20 | .00 | 61.209 | 21 | .00 | 90.581 | 20 | .00 | 77.307 | 19 | .00 | 61.119 | 18 | .00 |
| 14 | 66.695 | 20 | .00 | 75.971 | 21 | .00 | 56.172 | 20 | .00 | 89.125 | 19 | .00 | 57.163 | 18 | .00 |
| 15 | 54.846 | 23 | .00 | 87.684 | 22 | .00 | 56.824 | 22 | .00 | 66.017 | 22 | .00 | 40.928 | 22 | .01 |
| 16 | 87.640 | 23 | .00 | 99.367 | 24 | .00 | 80.408 | 23 | .00 | 56.694 | 23 | .00 | 69.568 | 23 | .00 |
| 17 | 85.888 | 23 | .00 | 42.308 | 22 | .01 | 64.091 | 22 | .00 | 43.308 | 21 | .00 | 45.631 | 21 | .00 |
| 18 | 108.753 | 23 | .00 | 114.282 | 24 | .00 | 135.375 | 23 | .00 | 118.059 | 23 | .00 | 180.510 | 23 | .00 |
| 19 | 135.323 | 23 | .00 | 187.295 | 24 | .00 | 153.039 | 23 | .00 | 171.377 | 23 | .00 | 177.980 | 23 | .00 |
| 20 | 69.049 | 23 | .00 | 57.552 | 23 | .00 | 105.441 | 22 | .00 | 48.646 | 21 | .00 | 83.070 | 22 | .00 |
| 21 | 252.141 | 22 | .00 | 287.101 | 21 | .00 | 212.003 | 22 | .00 | 318.718 | 21 | .00 | 252.603 | 21 | .00 |
| Total | 2336.055 | 456 | .00 | 2306.409 | 459 | .00 | 2063.117 | 441 | .00 | 2178.400 | 426 | .00 | 2361.751 | 431 | .00 |

*Note:* * denotes observed and expected frequencies are not statistically different (α > .01)

Table 4.32 Con’t

*Item and Scale Level Chi-Square Fit Statistics for Each Leadership Sample: PCM*

| | Sample 6 | | | Sample 7 | | | Sample 8 | | | Sample 9 | | | Sample 10 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Item | $\chi^2$ | *df* | *p* | $\chi^2$ | *df* | *p* | $\chi^2$ | *df* | *p* | $\chi^2$ | *df* | *p* | $\chi^2$ | *df* | *p* |
| 1 | 32.211 | 22 | .07* | 48.193 | 21 | .00 | 43.782 | 22 | .00 | 28.853 | 22 | .14* | 59.008 | 22 | .00 |
| 2 | 169.632 | 17 | .00 | 271.451 | 17 | .00 | 195.103 | 18 | .00 | 208.211 | 18 | .00 | 257.876 | 17 | .00 |
| 3 | 86.637 | 21 | .00 | 105.259 | 19 | .00 | 109.265 | 21 | .00 | 71.565 | 21 | .00 | 129.330 | 21 | .00 |
| 4 | 62.048 | 22 | .00 | 41.981 | 21 | .00 | 69.241 | 21 | .00 | 68.818 | 22 | .00 | 55.399 | 22 | .00 |
| 5 | 63.421 | 25 | .00 | 79.400 | 22 | .00 | 62.327 | 21 | .00 | 49.842 | 23 | .00 | 42.423 | 23 | .01 |
| 6 | 57.590 | 22 | .00 | 77.447 | 20 | .00 | 98.234 | 22 | .00 | 87.238 | 22 | .00 | 66.188 | 22 | .00 |
| 7 | 65.470 | 19 | .00 | 107.411 | 19 | .00 | 136.248 | 20 | .00 | 130.263 | 19 | .00 | 113.652 | 20 | .00 |
| 8 | 168.413 | 22 | .00 | 171.917 | 21 | .00 | 171.818 | 22 | .00 | 196.091 | 22 | .00 | 205.305 | 22 | .00 |
| 9 | 83.463 | 23 | .00 | 79.934 | 23 | .00 | 95.494 | 24 | .00 | 106.355 | 23 | .00 | 102.079 | 24 | .00 |
| 10 | 75.633 | 21 | .00 | 102.960 | 21 | .00 | 69.435 | 21 | .00 | 93.579 | 21 | .00 | 90.926 | 23 | .00 |
| 11 | 116.479 | 18 | .00 | 210.734 | 18 | .00 | 172.475 | 18 | .00 | 136.158 | 20 | .00 | 227.998 | 18 | .00 |
| 12 | 72.784 | 21 | .00 | 77.835 | 19 | .00 | 107.959 | 20 | .00 | 84.076 | 20 | .00 | 130.018 | 19 | .00 |
| 13 | 69.909 | 21 | .00 | 80.554 | 19 | .00 | 106.821 | 21 | .00 | 81.388 | 21 | .00 | 104.790 | 21 | .00 |
| 14 | 64.994 | 21 | .00 | 78.290 | 19 | .00 | 88.711 | 21 | .00 | 59.805 | 21 | .00 | 108.787 | 21 | .00 |
| 15 | 60.268 | 22 | .00 | 68.363 | 23 | .00 | 55.803 | 22 | .00 | 64.432 | 22 | .00 | 79.483 | 22 | .00 |
| 16 | 73.375 | 23 | .00 | 54.274 | 23 | .00 | 92.628 | 23 | .00 | 69.026 | 23 | .00 | 67.423 | 23 | .00 |
| 17 | 41.615 | 22 | .01 | 67.532 | 22 | .00 | 54.462 | 22 | .00 | 51.341 | 22 | .00 | 57.110 | 22 | .00 |
| 18 | 124.271 | 23 | .00 | 140.822 | 23 | .00 | 125.929 | 23 | .00 | 137.039 | 23 | .00 | 130.573 | 23 | .00 |
| 19 | 197.342 | 23 | .00 | 213.933 | 23 | .00 | 160.255 | 23 | .00 | 192.287 | 23 | .00 | 158.397 | 24 | .00 |
| 20 | 48.881 | 23 | .00 | 85.244 | 22 | .00 | 81.378 | 22 | .00 | 53.423 | 23 | .00 | 128.005 | 23 | .00 |
| 21 | 206.831 | 21 | .00 | 283.086 | 21 | .00 | 201.157 | 22 | .00 | 268.308 | 21 | .00 | 236.876 | 21 | .00 |
| Total | 1941.265 | 452 | .00 | 2446.619 | 436 | .00 | 2298.526 | 449 | .00 | 2238.100 | 452 | .00 | 2551.646 | 453 | .00 |

*Note:* * denotes observed and expected frequencies are not statistically different (α > .01)

Table 4.33

*Item and Scale Level Chi-Square Fit statistics for each Leadership Sample: GPCM*

| | Sample 1 | | | Sample 2 | | | Sample 3 | | | Sample 4 | | | Sample 5 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Item | $\chi^2$ | *df* | *p* | $\chi^2$ | *df* | *p* | $\chi^2$ | *df* | *p* | $\chi^2$ | *df* | *p* | $\chi^2$ | *df* | *p* |
| 1 | 48.287 | 23 | .00 | 43.948 | 22 | .00 | 49.120 | 23 | .00 | 49.864 | 23 | .00 | 68.686 | 23 | .00 |
| 2 | 142.438 | 21 | .00 | 175.938 | 24 | .00 | 100.649 | 23 | .00 | 119.481 | 23 | .00 | 157.388 | 22 | .00 |
| 3 | 58.051 | 18 | .00 | 68.461 | 18 | .00 | 51.355 | 18 | .00 | 50.362 | 18 | .00 | 44.824 | 18 | .00 |
| 4 | 55.977 | 23 | .00 | 37.679 | 21 | .01* | 72.570 | 23 | .00 | 63.586 | 23 | .00 | 48.974 | 21 | .00 |
| 5 | 55.614 | 23 | .00 | 80.899 | 21 | .00 | 57.889 | 22 | .00 | 76.554 | 22 | .00 | 89.497 | 21 | .00 |
| 6 | 56.572 | 19 | .00 | 51.791 | 22 | .00 | 54.310 | 22 | .00 | 62.888 | 21 | .00 | 66.621 | 19 | .00 |
| 7 | 71.594 | 22 | .00 | 62.833 | 22 | .00 | 77.834 | 20 | .00 | 95.764 | 20 | .00 | 118.640 | 23 | .00 |
| 8 | 48.868 | 18 | .00 | 65.262 | 17 | .00 | 45.640 | 17 | .00 | 43.595 | 17 | .00 | 61.372 | 17 | .00 |
| 9 | 76.961 | 24 | .00 | 59.212 | 20 | .00 | 38.721 | 21 | .01* | 74.084 | 21 | .00 | 46.371 | 21 | .00 |
| 10 | 40.608 | 21 | .01 | 40.363 | 19 | .00 | 49.939 | 21 | .00 | 80.750 | 19 | .00 | 44.304 | 21 | .00 |
| 11 | 226.643 | 21 | .00 | 168.727 | 20 | .00 | 138.554 | 21 | .00 | 158.347 | 19 | .00 | 175.720 | 19 | .00 |
| 12 | 105.190 | 20 | .00 | 64.873 | 20 | .00 | 97.642 | 18 | .00 | 105.507 | 18 | .00 | 112.347 | 20 | .00 |
| 13 | 84.642 | 20 | .00 | 80.771 | 21 | .00 | 99.675 | 22 | .00 | 91.149 | 19 | .00 | 113.146 | 22 | .00 |
| 14 | 103.623 | 22 | .00 | 73.151 | 19 | .00 | 62.377 | 20 | .00 | 78.380 | 20 | .00 | 76.102 | 20 | .00 |
| 15 | 61.211 | 23 | .00 | 102.286 | 22 | .00 | 60.218 | 22 | .00 | 61.537 | 22 | .00 | 43.906 | 22 | .00 |
| 16 | 98.039 | 23 | .00 | 60.786 | 20 | .00 | 77.610 | 23 | .00 | 61.449 | 22 | .00 | 54.124 | 21 | .00 |
| 17 | 98.927 | 23 | .00 | 63.063 | 22 | .00 | 75.573 | 23 | .00 | 48.799 | 22 | .00 | 55.036 | 21 | .00 |
| 18 | 51.107 | 22 | .00 | 35.702 | 20 | .01* | 55.510 | 20 | .00 | 62.037 | 20 | .00 | 52.787 | 18 | .00 |
| 19 | 40.550 | 21 | .01 | 65.295 | 19 | .00 | 43.690 | 20 | .00 | 44.551 | 18 | .00 | 43.565 | 18 | .00 |
| 20 | 78.642 | 25 | .00 | 96.780 | 24 | .00 | 108.856 | 24 | .00 | 73.190 | 23 | .00 | 109.790 | 23 | .00 |
| 21 | 509.716 | 25 | .00 | 673.659 | 27 | .00 | 457.924 | 26 | .00 | 573.654 | 25 | .00 | 604.407 | 26 | .00 |
| Total | 2113.258 | 457 | .00 | 2171.479 | 440 | .00 | 1875.653 | 449 | .00 | 2075.527 | 435 | .00 | 2187.608 | 436 | .00 |

*Note:* * denotes observed and expected frequencies are not statistically different (α > .01)

Table 4.33 Con’t

*Item and Scale level chi-square fit statistics for each Leadership Sample: GPCM*

|  | Sample 6 |  |  | Sample 7 |  |  | Sample 8 |  |  | Sample 9 |  |  | Sample 10 |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Item | $\chi^2$ | *df* | *p* | $\chi^2$ | *df* | *p* | $\chi^2$ | *df* | *p* | $\chi^2$ | *df* | *p* | $\chi^2$ | *df* | *p* |
| 1 | 49.826 | 23 | .00 | 58.551 | 23 | .00 | 49.578 | 22 | .00 | 45.602 | 23 | .00 | 56.311 | 22 | .00 |
| 2 | 106.662 | 23 | .00 | 134.946 | 23 | .00 | 152.163 | 24 | .00 | 158.505 | 23 | .00 | 177.507 | 22 | .00 |
| 3 | 54.355 | 21 | .00 | 54.098 | 18 | .00 | 53.409 | 18 | .00 | 33.428 | 20 | .03* | 62.736 | 19 | .00 |
| 4 | 71.004 | 24 | .00 | 47.489 | 21 | .00 | 79.588 | 23 | .00 | 71.506 | 23 | .00 | 74.857 | 23 | .00 |
| 5 | 81.141 | 25 | .00 | 90.243 | 21 | .00 | 90.971 | 23 | .00 | 62.720 | 24 | .00 | 71.572 | 25 | .00 |
| 6 | 45.938 | 21 | .00 | 71.256 | 20 | .00 | 92.297 | 22 | .00 | 47.168 | 21 | .00 | 46.762 | 22 | .00 |
| 7 | 44.604 | 21 | .00 | 97.483 | 23 | .00 | 87.714 | 24 | .00 | 97.870 | 22 | .00 | 80.483 | 24 | .00 |
| 8 | 13.620 | 17 | .69* | 26.056 | 15 | .03* | 52.990 | 17 | .00 | 29.445 | 18 | .04* | 52.326 | 18 | .00 |
| 9 | 107.594 | 24 | .00 | 74.383 | 20 | .00 | 60.285 | 21 | .00 | 106.763 | 23 | .00 | 105.765 | 22 | .00 |
| 10 | 48.399 | 21 | .00 | 73.541 | 21 | .00 | 33.404 | 21 | .04* | 80.119 | 21 | .00 | 48.487 | 21 | .00 |
| 11 | 110.168 | 19 | .00 | 160.211 | 20 | .00 | 164.765 | 21 | .00 | 129.745 | 20 | .00 | 208.038 | 22 | .00 |
| 12 | 53.262 | 20 | .00 | 72.035 | 18 | .00 | 118.604 | 20 | .00 | 82.861 | 20 | .00 | 129.960 | 20 | .00 |
| 13 | 67.116 | 21 | .00 | 77.240 | 21 | .00 | 106.896 | 22 | .00 | 87.965 | 21 | .00 | 108.507 | 21 | .00 |
| 14 | 66.587 | 21 | .00 | 72.772 | 19 | .00 | 91.409 | 21 | .00 | 50.987 | 20 | .00 | 99.461 | 20 | .00 |
| 15 | 67.505 | 23 | .00 | 85.753 | 22 | .00 | 67.115 | 22 | .00 | 77.548 | 22 | .00 | 93.119 | 22 | .00 |
| 16 | 74.717 | 23 | .00 | 48.257 | 22 | .00 | 87.692 | 23 | .00 | 81.862 | 23 | .00 | 64.904 | 22 | .00 |
| 17 | 45.321 | 22 | .00 | 86.478 | 22 | .00 | 66.531 | 22 | .00 | 59.681 | 22 | .00 | 72.962 | 22 | .00 |
| 18 | 23.684 | 20 | .25* | 50.918 | 19 | .00 | 33.500 | 19 | .02* | 21.784 | 20 | .35* | 26.767 | 20 | .14* |
| 19 | 17.098 | 19 | .58* | 58.668 | 18 | .00 | 64.196 | 19 | .00 | 45.954 | 20 | .00 | 35.824 | 20 | .02* |
| 20 | 49.899 | 24 | .00 | 99.217 | 24 | .00 | 96.461 | 23 | .00 | 87.571 | 24 | .00 | 118.127 | 24 | .00 |
| 21 | 521.610 | 26 | .00 | 570.885 | 25 | .00 | 569.606 | 27 | .00 | 627.478 | 26 | .00 | 538.006 | 26 | .00 |
| Total | 1720.112 | 458 | .00 | 2110.480 | 435 | .00 | 2219.175 | 454 | .00 | 2086.563 | 456 | .00 | 2272.484 | 457 | .00 |

*Note:* * denotes observed and expected frequencies are not statistically different (α > .01)

Table 4.34 displays the chi-square distributed fit statistics ($G^2$) from the GGUM analyses. The same caution as in the Empowerment analyses must be exercised when interpreting the fit results produced from the GGUM, particularly for those items associated with few fit groups. Few fit groups logically are associated with fewer degrees of freedom, which will influence the chi-square statistic. Clearly more items were found to fit better within the GGUM analyses; however, determination of fit based on these results alone would be inappropriate because good fit could be a function of the reduced degrees of freedom. Nonetheless, some patterns did emerge from the GGUM fit results, namely the consistent lack of fit of items 2, 11, and 21 across the 10 samples. According to the GGUM analyses items 2 and 21 had the two lowest discrimination estimates and items 11 and 21 had the most modest item location estimates. Items 2 and 11 ask respondents to rate their level of agreement with the statements: “The faculty are committed to helping every student learn” and “Teachers are held to high professional standards for delivering instruction.” Item 21 reads “Overall, the school leadership in my school is effective.”

Table 4.34

*Item and Scale Level Chi-Square Fit Statistics for Each Leadership Sample: GGUM*

| | Sample 1 | | | | Sample 2 | | | | Sample 3 | | | | Sample 4 | | | | Sample 5 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Item | $\chi^2$ | *df* | *p* | *fit grps.* | $\chi^2$ | *df* | *p* | *fit grps.* | $\chi^2$ | *df* | *p* | *fit grps.* | $\chi^2$ | *df* | *p* | *fit grps.* | $\chi^2$ | *df* | *p* | *fit grps.* |
| 1 | 28.274 | 16 | .029* | 4 | 17.701 | 16 | .342* | 4 | 14.133 | 16 | .589* | 4 | 29.098 | 20 | .086* | 5 | 30.655 | 20 | 0.060* | 5 |
| 2 | 44.135 | 16 | .000 | 4 | 59.698 | 20 | .000 | 5 | 38.878 | 16 | .001 | 4 | 46.235 | 16 | .000 | 4 | 53.189 | 16 | .000 | 4 |
| 3 | 8.981 | 8 | .343* | 2 | 15.731 | 8 | .046* | 2 | 17.744 | 12 | .124* | 3 | 21.561 | 12 | .043* | 3 | 6.252 | 8 | .619* | 2 |
| 4 | 25.280 | 16 | .065* | 4 | 9.516 | 16 | .891* | 4 | 47.160 | 24 | .003 | 6 | 36.794 | 24 | .046* | 6 | 27.683 | 16 | .035* | 4 |
| 5 | 15.924 | 20 | .721* | 5 | 33.108 | 20 | .033* | 5 | 29.830 | 20 | .073* | 5 | 28.884 | 20 | .090* | 5 | 37.487 | 20 | .010* | 5 |
| 6 | 9.102 | 12 | .694* | 3 | 14.976 | 12 | .243* | 3 | 27.172 | 16 | .040* | 4 | 19.856 | 16 | .227* | 4 | 11.054 | 12 | .524* | 3 |
| 7 | 29.066 | 16 | .023* | 4 | 30.885 | 16 | .014* | 4 | 15.894 | 12 | .196* | 3 | 33.691 | 12 | .001 | 3 | 47.832 | 16 | .000 | 4 |
| 8 | 0.497 | 4 | .973* | 1 | 0.228 | 4 | .994* | 1 | 20.802 | 8 | .008 | 2 | 0.411 | 4 | .982* | 1 | 0.323 | 4 | .988* | 1 |
| 9 | 26.592 | 12 | .008* | 3 | 14.516 | 8 | .069* | 2 | 12.436 | 12 | .411* | 3 | 14.765 | 12 | .255* | 3 | 8.086 | 12 | .778* | 3 |
| 10 | 6.257 | 8 | .618* | 2 | 8.254 | 8 | .409* | 2 | 2.823 | 8 | .945* | 2 | 12.748 | 8 | .121* | 2 | 10.891 | 8 | .208* | 2 |
| 11 | 99.036 | 16 | .000 | 4 | 73.102 | 16 | .000 | 4 | 91.668 | 16 | .000 | 4 | 88.370 | 12 | .000 | 3 | 65.116 | 16 | .000 | 4 |
| 12 | 16.737 | 12 | .160* | 3 | 9.266 | 12 | .680* | 3 | 18.727 | 12 | .095* | 3 | 19.485 | 12 | .078* | 3 | 36.466 | 16 | .003 | 4 |
| 13 | 10.100 | 12 | .607* | 3 | 19.887 | 12 | .069* | 3 | 26.471 | 16 | .048* | 4 | 26.945 | 12 | .008 | 3 | 10.534 | 12 | .569* | 3 |
| 14 | 18.165 | 12 | .111* | 3 | 20.599 | 12 | .057* | 3 | 23.924 | 12 | .028* | 3 | 29.731 | 12 | .003 | 3 | 9.190 | 12 | .687* | 3 |
| 15 | 18.526 | 16 | .294* | 4 | 4.792 | 8 | .780* | 2 | 20.548 | 16 | .197* | 4 | 36.712 | 16 | .002 | 4 | 5.421 | 8 | .712* | 2 |
| 16 | 23.799 | 12 | .022* | 3 | 9.575 | 8 | .296* | 2 | 24.408 | 12 | .018* | 3 | 12.288 | 12 | .423* | 3 | 16.930 | 8 | .031* | 2 |
| 17 | 32.260 | 16 | .009 | 4 | 20.319 | 16 | .206* | 4 | 15.003 | 12 | .241* | 3 | 4.302 | 12 | .977* | 3 | 6.927 | 8 | .545* | 2 |
| 18 | 0.543 | 4 | .969* | 1 | 0.231 | 4 | .994* | 1 | 0.243 | 4 | .993* | 1 | 0.230 | 4 | .994* | 1 | 0.280 | 4 | .991* | 1 |
| 19 | 0.732 | 4 | .947* | 1 | 0.384 | 4 | .984* | 1 | 0.448 | 4 | .978* | 1 | 0.439 | 4 | .979* | 1 | 0.563 | 4 | .967* | 1 |
| 20 | 33.955 | 20 | .026* | 5 | 36.687 | 20 | .013* | 5 | 48.360 | 20 | .000 | 5 | 27.343 | 20 | .126* | 5 | 30.998 | 16 | .014* | 4 |
| 21 | 698.81 | 24 | .00 | 6 | 1020.80 | 32 | .00 | 8 | 772.97 | 32 | .00 | 8 | 898.31 | 32 | .00 | 8 | 945.58 | 32 | .00 | 8 |
| Total | 1146.78 | 69 | .00 | | 1420.26 | 68 | .00 | | 1269.64 | 75 | .00 | | 1388.20 | 73 | .00 | | 1361.46 | 67 | .00 | |

*Note:* * denotes observed and expected frequencies are not statistically different (α > .01)

Table 4.34 Con’t

*Item and Scale Level Chi-Square Fit Statistics for Each Leadership Sample: GGUM*

| | Sample 6 | | | | Sample 7 | | | | Sample 8 | | | | Sample 9 | | | | Sample 10 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Item | $\chi^2$ | *df* | *p* | *fit grps.* | $\chi^2$ | *df* | *p* | *fit grps.* | $\chi^2$ | *df* | *p* | *fit grps.* | $\chi^2$ | *df* | *p* | *fit grps.* | $\chi^2$ | *df* | *p* | *fit grps.* |
| 1 | 24.588 | 20 | .218* | 5 | 24.268 | 20 | .231* | 5 | 25.008 | 16 | .070* | 4 | 9.200 | 16 | .905* | 4 | 5.640 | 12 | .933* | 3 |
| 2 | 36.946 | 16 | .002 | 4 | 53.822 | 16 | .000 | 4 | 54.414 | 20 | .000 | 5 | 90.737 | 16 | .000 | 4 | 65.946 | 16 | .000 | 4 |
| 3 | 10.739 | 12 | .551* | 3 | 0.848 | 4 | .932* | 1 | 11.710 | 12 | .469* | 3 | 10.484 | 12 | .574* | 3 | 16.585 | 8 | .035* | 2 |
| 4 | 26.387 | 24 | .334* | 6 | 9.943 | 16 | .869* | 4 | 43.684 | 24 | .008 | 6 | 21.771 | 16 | .151* | 4 | 26.963 | 16 | .042* | 4 |
| 5 | 29.250 | 20 | .083* | 5 | 28.125 | 16 | .031* | 4 | 35.868 | 20 | .016* | 5 | 23.420 | 16 | .103 | 4 | 35.825 | 20 | .016* | 5 |
| 6 | 27.239 | 16 | .039* | 4 | 12.225 | 12 | .428* | 3 | 29.439 | 16 | .021* | 4 | 6.532 | 8 | .588* | 2 | 26.079 | 16 | .053* | 4 |
| 7 | 27.670 | 12 | .006 | 3 | 25.452 | 16 | .062* | 4 | 17.979 | 16 | .325* | 4 | 32.842 | 12 | .001 | 3 | 39.082 | 16 | .001 | 4 |
| 8 | 0.337 | 4 | .987* | 1 | 0.408 | 4 | .982* | 1 | 12.928 | 8 | .114* | 2 | 0.301 | 4 | .989* | 1 | 0.514 | 4 | .972* | 1 |
| 9 | 14.263 | 12 | .284* | 3 | 14.943 | 12 | .245* | 3 | 13.966 | 12 | .303* | 3 | 33.843 | 12 | .001 | 3 | 14.277 | 12 | .283* | 3 |
| 10 | 9.986 | 8 | .266* | 2 | 26.377 | 12 | .009 | 3 | 7.599 | 8 | .474* | 2 | 16.948 | 8 | .031* | 2 | 8.101 | 8 | .424* | 2 |
| 11 | 63.200 | 12 | .000 | 3 | 122.848 | 16 | .000 | 4 | 37.123 | 16 | .002 | 4 | 51.791 | 12 | .000 | 3 | 137.341 | 16 | .000 | 4 |
| 12 | 26.527 | 12 | .009 | 3 | 13.574 | 12 | .329* | 3 | 17.165 | 12 | .144* | 3 | 10.499 | 12 | .572* | 3 | 21.909 | 12 | .038* | 3 |
| 13 | 29.817 | 12 | .003 | 3 | 17.149 | 12 | .144* | 3 | 19.102 | 12 | .086* | 3 | 17.565 | 12 | .129* | 3 | 26.503 | 12 | .009 | 3 |
| 14 | 35.228 | 12 | .000 | 3 | 19.532 | 12 | .076* | 3 | 16.447 | 12 | .172* | 3 | 13.058 | 12 | .364* | 3 | 24.072 | 12 | .020* | 3 |
| 15 | 23.135 | 16 | .110* | 4 | 13.989 | 12 | .301* | 3 | 11.871 | 12 | .456* | 3 | 16.386 | 16 | .426* | 4 | 23.700 | 8 | .003 | 2 |
| 16 | 18.349 | 12 | .106* | 3 | 5.815 | 8 | .668* | 2 | 19.250 | 12 | .083* | 3 | 7.508 | 12 | .822* | 3 | 22.038 | 8 | .005 | 2 |
| 17 | 15.897 | 12 | .196* | 3 | 22.640 | 12 | .031* | 3 | 4.957 | 8 | .762* | 2 | 21.835 | 12 | .039* | 3 | 20.903 | 8 | .007 | 2 |
| 18 | 0.291 | 4 | .990* | 1 | 0.400 | 4 | .983* | 1 | 0.226 | 4 | .994* | 1 | 0.219 | 4 | .994* | 1 | 0.454 | 4 | .978* | 1 |
| 19 | 0.580 | 4 | .965* | 1 | 0.605 | 4 | .963* | 1 | 0.430 | 4 | .98* | 1 | 0.441 | 4 | .979* | 1 | 0.680 | 4 | .954* | 1 |
| 20 | 25.251 | 20 | .192* | 5 | 33.455 | 20 | .030* | 5 | 36.547 | 12 | .000 | 3 | 19.896 | 20 | .465* | 5 | 80.506 | 20 | .000 | 5 |
| 21 | 833.62 | 28 | .00 | 7 | 957.81 | 32 | .00 | 8 | 886.45 | 32 | .00 | 8 | 957.67 | 32 | .00 | 8 | 803.09 | 28 | .00 | 7 |
| Total | 1279.30 | 72 | .00 | | 1404.22 | 68 | .00 | | 1302.16 | 72 | .00 | | 1362.94 | 67 | .00 | | 1400.21 | 65 | .00 | |

*Note:* * denotes observed and expected frequencies are not statistically different (α > .01)

Another approach used for assessing model fit was the calculation of AIC and BIC statistics with the application of the IRT models to the Leadership data. Table 4.35 displays those fit statistics estimated by the PCM, GPCM, and GGUM models. These results mimic those of the Empowerment analyses in that the GGUM was associated with much smaller AIC and BIC values than the GPCM, and the differences between the GPCM and PCM solutions were small. Based on these results, the GGUM exhibited the best fit to the Leadership data, while GPCM fit the data better than the PCM. In sum, analyses conducted to answer the fourth research question reveal that, across the three IRT models, the items generally do not fit well and neither do the models according to the chi-square statistics. According to the AIC and BIC criteria, the GGUM fit relatively better than the GPCM, though no criterion exists to measure 'how much' better.

Table 4.35

*AIC and BIC Results for Each Leadership Sample from PCM, GPCM, and GGUM models*

| Sample | PCM | | GPCM | | GGUM | |
|---|---|---|---|---|---|---|
| | AIC | BIC | AIC | BIC | AIC | BIC |
| 1 | 85275.831 | 85863.926 | 84442.372 | 85148.086 | 57900.66 | 59563.71 |
| 2 | 85692.078 | 86280.173 | 84611.837 | 85317.551 | 59563.71 | 59832.398 |
| 3 | 85566.826 | 86154.921 | 84840.242 | 85545.956 | 58169.602 | 60046.716 |
| 4 | 84823.419 | 85411.514 | 84061.782 | 84767.496 | 59832.398 | 59253.97 |
| 5 | 83862.589 | 84450.684 | 82839.962 | 83545.676 | 56370.513 | 58033.562 |
| 6 | 84556.071 | 85144.166 | 83715.5 | 84421.214 | 57243.101 | 58905.897 |
| 7 | 84715.443 | 85303.538 | 83672.425 | 84378.139 | 57205.395 | 58868.57 |
| 8 | 85700.59 | 86288.685 | 84749.879 | 85455.593 | 58268.865 | 59932.04 |
| 9 | 84171.773 | 84759.868 | 83244.731 | 83950.445 | 56781.012 | 58444.061 |
| 10 | 83506.736 | 84094.831 | 82572.435 | 83278.149 | 56122.599 | 57785.522 |

*Person Locations*

The focus of the second research question had to do with the location of the sample on the latent trait and the ordering of respondents on the latent trait across IRT models. IRT calibrations were conducted for the PCM, GPCM, and GGUM on the 10 Leadership samples and person parameters are provided in this section. Rank-order correlations and scatterplots of the person parameters are presented to address research question two. The relationship between the rank ordering of the person parameter trait estimates produced by the PCM, GPCM, and CFA scaling methods and the GGUM was examined with simple Kendall Tau-*b* correlations. Kendall's Tau-*b* correlations of the person trait estimates across the 10 Leadership samples for each pair of scaling methods are presented in Table 4.36, along with the average correlation across samples.

Table 4.36

*Kendall's Tau-b Correlations among Person Trait Estimates by Sample and Scaling Method*

| Sample | PCM,<br>GGUM | GPCM,<br>GGUM | PCM,<br>GPCM | CFA,<br>PCM | CFA,<br>GPCM | CFA,<br>GGUM |
|---|---|---|---|---|---|---|
| 1 | .973 | .994 | .974 | .942 | .955 | .956 |
| 2 | .969 | .999 | .970 | .941 | .959 | .959 |
| 3 | .971 | .999 | .972 | .940 | .955 | .955 |
| 4 | .972 | .999 | .973 | .942 | .956 | .956 |
| 5 | .963 | .999 | .964 | .929 | .951 | .952 |
| 6 | .970 | .999 | .971 | .939 | .952 | .952 |
| 7 | .966 | .999 | .966 | .935 | .955 | .956 |
| 8 | .968 | .999 | .969 | .938 | .955 | .955 |
| 9 | .970 | .999 | .971 | .965 | .982 | .982 |
| 10 | .971 | .999 | .971 | .940 | .957 | .957 |
| Mean | .969 | .998 | .970 | .941 | .958 | .958 |

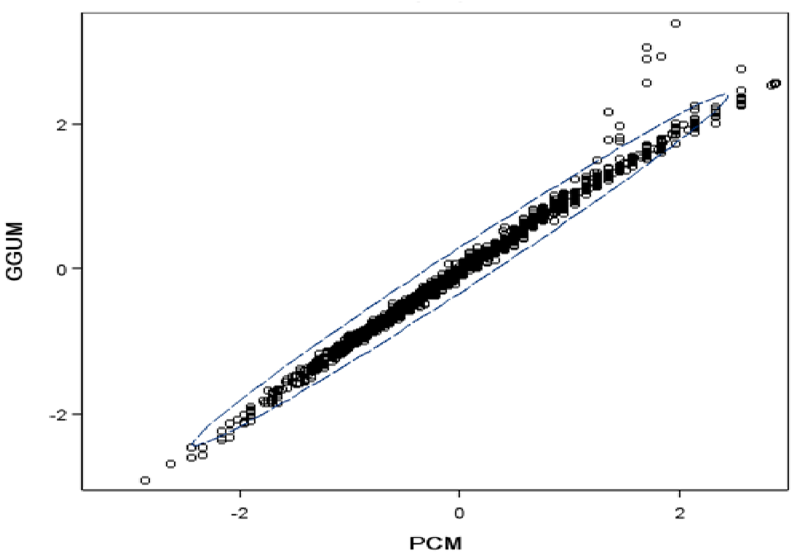
Two trends visible in Table 4.36 were also apparent in the Empowerment analyses. First, the greatest measure of association between the trait estimates was found between the GPCM and GGUM models, indicating nearly identical ordering of person estimates. Second, the smallest correlations were consistently found between the CFA and PCM models. One difference between the Leadership and Empowerment results is found in the strength of association in that all correlations (and average correlations) were lower in the Empowerment analyses. For example, the smallest mean correlation in the Empowerment analyses between the PCM and CFA was lower (Tau = .873) than the average correlation between the same scaling methods in the Leadership analysis (Tau = .941). The inconsistency in the rank order of person estimates between the PCM and CFA models is less pronounced in the Leadership analyses. Finally, all correlations in each of the sample, for all combinations of scaling methods were statistically significant (*p* < .01).

Graphical representations of the relationship between the trait estimates are found in the form of scatterplots in Figures 4.108, 4.109, and 4.110. These plots show the association of trait estimates between the GGUM and the PCM, GPCM, and CFA models, respectively, from sample 1.

The trait estimates produced by the PCM and GGUM were very similar with the exception of some outliers at the upper end of the distribution. Those cases were estimated to have higher GGUM values than PCM trait values. The relationship between the GPCM and GGUM trait estimates shown in Figure 4.109 was nearly linear, with the exception of outliers at the upper end of the trait.

Figure 4.108

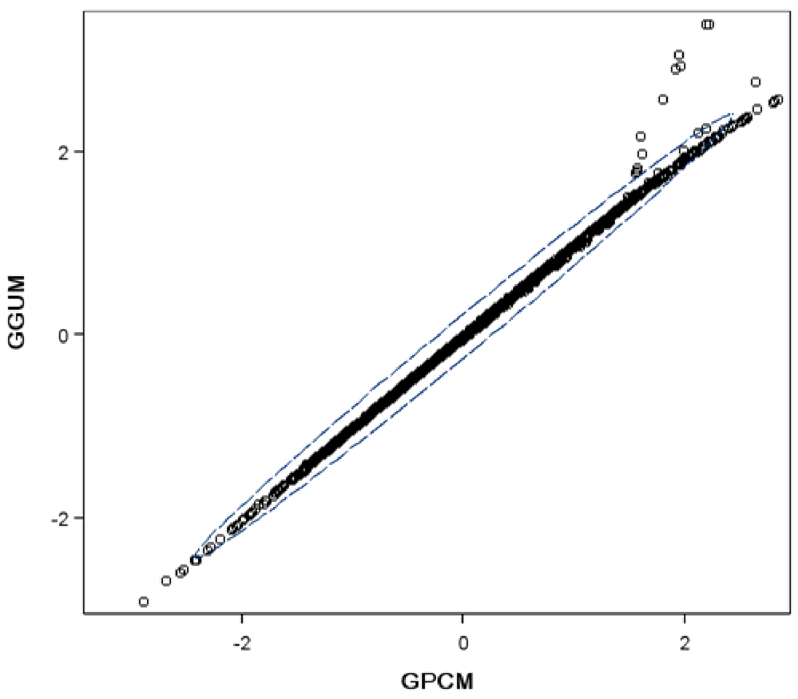
*Scatterplot of Trait Estimates for PCM and GGUM models: Sample 1 Leadership Scale*



*Note. Dashed lines denotes 95% prediction confidence ellipse*

Figure 4.109

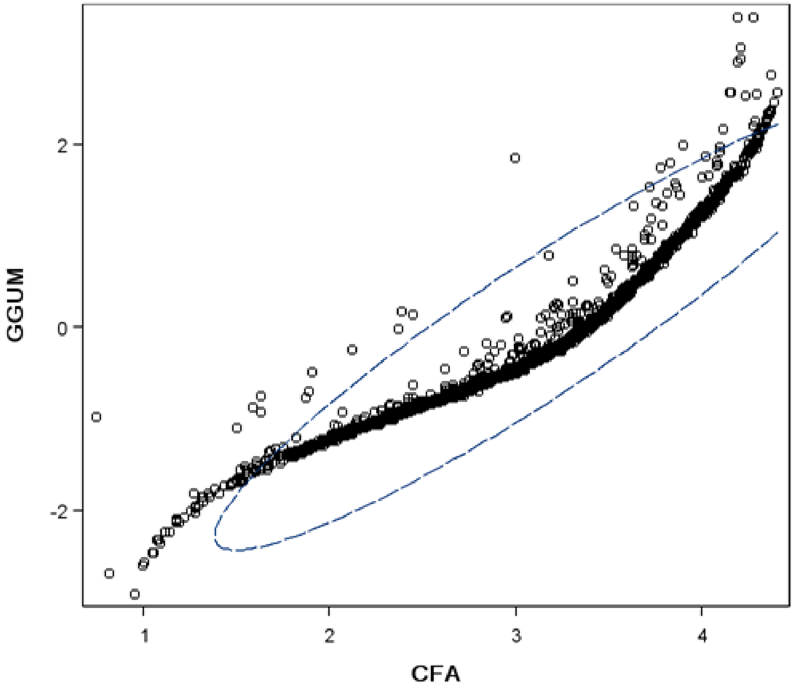
*Scatterplot of Trait Estimates for GPCM and GGUM models: Sample 1 Leadership Scale*



*Note.* Dashed line denotes 95% prediction confidence interval

Figure 4.110

*Scatterplot of Trait Estimates for CFA and GGUM models: Sample 1 Leadership Scale*



*Note. Dashed line denotes 95% prediction confidence ellipse*

Finally, the nonlinear relationship between the CFA and GGUM trait estimates are shown in Figure 4.110. Towards the middle of the distribution the CFA model generates higher estimates than the GGUM, and closer to the positive end of the trait, the GGUM appears to yield higher trait estimates than the CFA model. Outliers are present at both ends of the latent trait, and even along most of the continuum. Across the three (PCM, GPCM, and CFA) models, the CFA functions most inconsistently with the GGUM.

Another approach used to examine the estimated trait distributions and to further detect discrepancies in trait estimates between models, was the calculation 5 X 5 cross-tabulation tables. The procedures used were identical to those of the Empowerment analysis. Results are presented again only for the first Leadership sample. The frequency of respondents within given quintiles for each pair of models with the GGUM are presented in Table 4.37 for the PCM, Table 4.38 for the GPCM, and Table 4.39 for the CFA model.

Table 4.37

*Cross Tabulation Table of GGUM and PCM Quintiles: Sample 1 Leadership Scale*

| | | | | PCM | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | Total | Statistic | Value | ASE |
| GGUM | 1 | 386 | 13 | 0 | 0 | 0 | 399 | Tau-*b* | .971 | .003 |
| | 2 | 11 | 374 | 15 | 0 | 0 | 400 | Tau-*c* | .971 | .003 |
| | 3 | 0 | 13 | 377 | 9 | 0 | 399 | Pearson | .985 | .001 |
| | 4 | 0 | 0 | 44 | 349 | 7 | 400 | Spearman | .985 | .001 |
| | 5 | 0 | 0 | 0 | 6 | 393 | 399 | | | |
| | Total | 397 | 400 | 436 | 364 | 400 | 1997 | | | |

*Notes:* ASE = Asymptotic Standard Error; Tau-*b* = Kendall's Tau-*b;* Tau-*c* = Stuart's Tau-*c*

The frequencies within each quintile by model reveal that, across the five trait categories, the GGUM consistently produced higher estimates within the fourth quintile. When the other models estimated higher person traits, those also fell into the fourth quintile, with the exception of the CFA model.

Table 4.38

*Cross Tabulation Table of GGUM and GPCM Quintiles: Sample 1 Leadership Scale*

| | | | | | GPCM | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | Total | Statistic | Value | ASE |
| GGUM | 1 | 396 | 3 | 0 | 0 | 0 | 399 | Tau-b | .995 | .001 |
| | 2 | 1 | 397 | 2 | 0 | 0 | 400 | Tau-c | .995 | .001 |
| | 3 | 0 | 0 | 393 | 6 | 0 | 399 | Pearson | .998 | .001 |
| | 4 | 0 | 0 | 5 | 393 | 2 | 400 | Spearman | .998 | .001 |
| | 5 | 0 | 0 | 0 | 1 | 398 | 399 | | | |
| | Total | 397 | 400 | 400 | 400 | 400 | 1997 | | | |

*Notes:* ASE = Asymptotic Standard Error; Tau-*b* = Kendall's Tau-*b*; Tau-*c* = Stuart's Tau-*c*

For example, the GGUM estimated 44 respondents within the fourth quintile, while the PCM categorized those 44 people into the third quintile. The CFA categorized 30 people in the fourth quintile, while the GGUM estimated those 30 to fall into the third quintile. Further, as seen in Table 4.37, the GGUM resulted in higher trait estimates at the higher end (i.e., 4th and 5th quintiles) of the trait scale.

Table 4.39

*Cross Tabulation Table of GGUM and CFA Quintiles: Sample 1 Leadership Scale*

| | | | | CFA | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | Total | Statistic | Value | ASE |
| | 0 | 383 | 16 | 0 | 0 | 0 | 399 | Tau-*b* | .956 | .004 |
| GGUM | 1 | 10 | 365 | 25 | 0 | 0 | 400 | Tau-*c* | .956 | .004 |
| | 2 | 2 | 16 | 351 | 30 | 0 | 399 | Pearson | .975 | .003 |
| | 3 | 2 | 2 | 23 | 357 | 16 | 400 | Spearman | .975 | .003 |
| | 4 | 0 | 1 | 1 | 13 | 384 | 399 | | | |
| | Total | 397 | 400 | 400 | 400 | 400 | 1997 | | | |

*Notes*: ASE = Asymptotic Standard Error; Tau-*b* = Kendall's Tau-*b*; Tau-*c* = Stuart's Tau-*c*

### *Summary of Leadership Analyses*

Results reveal that the dimensionality assumptions of the cumulative models were met, and that assumption of unidimensionality within the context of unfolding models was not. The data do not appear to be of the unfolding type (i.e., responses do not unfold) based on the structure of the data. Further, the plot of pattern coefficients that resulted from the PCA with two components did not produce a circumplex-like (i.e., semi-circular) structure. Most of the fit indices produced by the CFA were supporting of excellent model fit and that a single factor structure was adequate in explaining the leadership data. Overall, these results support the assumptions of cumulative IRT models.

As for the item parameters, the PCM and GPCM performed similarly across all item parameters estimated. Both models indicated that most of the Leadership items were moderately easy to endorse. The GGUM analyses estimated all Leadership items to have extreme location estimates and were associated with very large standard errors. According to the GGUM, all items were clustered in a narrow and extreme region of the latent trait. The Leadership items are not extremely worded in either direction, therefore the extremity of item parameter estimates could be an outcome of relative homogeneity of attitudes of the sample. Because item location estimates and signs of those estimates are also associated with item content, according to the GGUM analyses, moderate and negative attitudes towards Leadership were not measured by these 21 items. The similarity across the three IRT models is that the 21 items are closely clustered in a small region of the latent trait. Therefore, much of the latent trait is not being measured by the 21 Leadership items. The category probability plot produced from the GGUM for item 21 showed characteristics of an unfolding item across the five response options. More moderate unfolding characteristics were seen in the category probability plots for items 7, and 11 through 17 for the Agree and Strongly Agree response options. The category probability plots produced from the PCM, GPCM, and GGUM for the remainder of the items are similar, implying that the three IRT models function similarly.

Across the 10 PCM analyses, almost none of the items fit the Leadership data well according to the item level chi-square fit statistics. The GPCM analyses displayed only few items that fit well, and both cumulative models fit statistically poorly at the scale level. The GGUM item level fit statistics reveal that many items fit the data, but with some items using as few as 1 or 2 or 3 fit groups to calculate the chi-square statistic, results are questionable. At the scale level, however, the GGUM also did not fit well across the 10 samples. The AIC and BIC measures of fit indicated that the GGUM fit the Leadership data better than the GPCM, and that the GPCM fit better than the PCM.

Finally, in examining person parameter estimates and distributions, the weakest relationship between the person trait parameter estimates was found between the CFA and PCM scaling methods. The model that exhibited the least amount of agreement with the GGUM was the CFA, while, just as in the Empowerment analyses, the GPCM and GGUM were essentially indistinguishable in person parameter estimates based on the rank order correlations and scatterplots.

# CHAPTER FIVE

# CONCLUSIONS AND DISCUSSION

The measurement of non-cognitive constructs such as attitudes, preference, opinion, and psychological constructs including personality has been undertaken by psychologists, market researchers, across a variety of fields, such as industrial/organizational psychology, personality psychology and even within the armed services. Examination of change in behavior and developmental/stage processes, and the relationship between attitudes and behaviors have certainly been of interest to developmental and educational psychologists. Traditionally, cumulative IRT models have been applied to non-cognitive data as such models were originally developed for the measurement of cognitive constructs like achievement and aptitude. Through the early contributions of Thurstone (1927, 1928), Coombs (1950, 1961), and more recently Andrich (1978, 1988, 1995, 1996,) Luo (1998, 2000), Roberts and colleagues (1996, 1998, 1999, 2000, 2002, 2003), and Chernyshenko, Stark, Drasgow and colleagues (2001, 2006, 2007), theory development underlying how people respond to attitudinal Likert-type items and the development of software to analyze such data have enabled the application of the relatively less familiar unfolding IRT models.

In the literature there are generally three perspectives from which applied researchers come when employing unfolding IRT models. The first, and that taken in this investigation, is for the purpose of informing non-cognitive scale construction. Closely related, and perhaps a natural result of improved or better informed scale construction is the perspective of improved accuracy and measurement precision, supporting the use of unfolding IRT models over cumulative IRT models. Third, unfolding models have been applied as a novel approach for examining and explaining the relationship between attitudes and behaviors (Andrich & Styles, 1998; Noel, 1999). This contemporary way of thinking about the attitude-behavior relationship has changed the way attitudes and behaviors are measured with the application of models that accommodate single-peaked functions to measure this association and the way in which behaviors change over time.

Unfolding IRT models have been shown not only to function as “viable alternatives” (Stark et al., 2006, p. 25) to cumulative models, but in some measurement situations, they are superior to traditionally used cumulative models in more accurately measuring non-cognitive latent constructs, and in developing scales for measuring such constructs (Roberts et al., 1999). These measurement situations include those in which an ideal point response process is responsible for producing the observed data. The problem with using the traditional cumulative IRT models to measure non-cognitive constructs (measured by Likert-type items) is that, theoretically, people use an ideal point response process when asked to rate their level of agreement with an item. The argument against the use of cumulative models in these situations is that there exists a discrepancy between the assumptions that underlie the data generated from the ideal point response process and those that underlie the model (that is, cumulative IRT models that assume a dominance response process).

The approach used in this investigation was to examine how unfolding IRT models function compared to traditional, cumulative IRT models. Because some details regarding the development of the NCTWCS were not available, careful and close examination of all the factors related to model assumptions, model fit, and characteristics of item and person parameters was undertaken. Whereas much of the applied research using unfolding IRT models has focused on attitudes of college students towards capital punishment and abortion, the intent of this investigation was to inform scale construction and analysis for a different construct and population – teachers' perception of and attitudes about school leadership (Leadership) and teacher empowerment (Empowerment).

In the following sections of this chapter, the limitations to this investigation. Following that, important findings are reviewed and interpreted. Potential reasons for the findings are proposed by scale (Empowerment and Leadership), and implications of the findings and suggestions for future research are suggested.

## Limitations

Six limitations of the methodology used in this investigation are apparent. First, the nesting of teachers within schools was not considered in this analysis. This is only a limitation because the entire sample could not be used. If the entire sample could have been used, nesting would not pose a threat, as the intent of the NCTWCS was to examine leadership practices, and make decisions at the school/principal level. Decisions were not made at the individual teacher level. Accommodation of the nested structure of the data could have been accounted for by employing hierarchical IRT models and hierarchical CFA models. Observations in hierarchical or nested data (i.e., teachers nested within schools) are dependent and tend to be more similar to each other in terms of the outcome variable (i.e., measure of Leadership or Empowerment) than they do to those in a different group (i.e., school). These dependencies result in problems when the focus is on the individual level observations that are nested within groups. Second, it is possible that bias could have been introduced into the estimates of as a result of survey non-response, as the average response rate, across *schools* was approximately 66%. It could be possible that non-response occurred for similar reasons among individuals who did not participate. If those individuals possess common attributes that directly relate to the construct being measured (attitudes toward teacher working conditions), then systematic non-response would not be captured in the analyses. Further, a sub-sample of the total population omitted could contribute to a relatively homogeneous effective sample. However, if the occurrences of non-responses were random, then bias would not be a concern. Given that the NCTWCS has only been administered three times (2002, 2004, and 2006), that the prior years' data were not available, and that individual teachers (and their characteristics) could not be tracked due to the anonymity and confidentiality built into the data collection procedures, it was not possible to use auxiliary data to examine or statistically adjust for non-response bias. A third key limitation in this investigation was that the “true” attitudes toward teacher Empowerment and Leadership were not known, therefore an absolute decision of a best or correct model in terms of measurement precision and accuracy could not be made.

A fourth limitation concerns model selection. This investigation used a common Rasch cumulative model (the PCM), a common cumulative IRT model for polytomous data (the GPCM), and an unfolding IRT model for polytomous data (the GGUM). All of these models are also parametric models. Other models could have been selected, especially for this kind of attitudinal data. Nonparametric cumulative IRT models were not used in this study, although their use in the field of cognitive and non-cognitive measurement is perfectly acceptable. Some researchers argue their superiority over parametric IRT models simply because the strict assumptions that underlie parametric IRT models are difficult to meet.

Competing theories in the context of psychological measurement include the idea that nonparametric cumulative IRT models are more appropriate and more efficient than parametric IRT models (Cliff et al., 1998; Collins et al., 2006; Meijer & Baneke, 2004; Nandakumar et al., 2002; Rabe-Hesketh & Skrondal, 2007). There are also proponents of the application of nonparametric unfolding IRT models to non-cognitive constructs including Cliff et al. (1988) and van Schuur (1984). The argument supporting nonparametric models is that these models provide more flexibility because less strict statistical assumptions are required. This claim is certainly valid, however, the purpose of the survey and use of results and future administrations must be determined prior to selection of analysis method. If the benefits of IRT are sought, then nonparametric IRT models may not allow for the greatest advantage as the latent space is not completely specified therefore inhibiting the satisfaction of the assumptions item independence and invariance. Complete specificity of the latent space refers to a measurement model that "completely specifies the relation between person location, stimulus location and choice probability" (Hoijtink, 1990, p. 642). However, if strict adherence to assumptions of IRT models cannot be met, or are believed to be violated to some degree, the use of nonparametric measurement models is likely more appropriate than parametric models. In his chapter on locally dependent conjunctive IRT models, Jannarone (1997) claims that "from formal and practical viewpoints, therefore, the local independence axiom stands in the way of interesting extensions to test theory and application" (p. 472). The argument for the use of nonparametric IRT models for analyzing non-cognitive data is warranted especially given the little empirical research on application of nonparametric unfolding IRT models to such data.

A fifth feature of this study that arguably poses limitations is that because multilevel IRT and SEM modeling were not employed, teacher-level trait estimates were not or could not be aggregated to the school level. School level measures would have been useful as the NCTWCS data are used to make decisions about schools, or the principals within schools, based on teacher level responses. If the teacher level trait estimates could have been aggregated to the school level in this investigation, then further comparisons could have been made between model results and the school level decisions that were made by the Governor’s office of North Carolina. This last point was also inhibited due to the fact that the entire sample could not have been used in the GGUM analyses due to software constraints, and thus for all analyses in this investigation.

Finally, in all the applied research that use unfolding IRT models, data are only dichotomous (agree/disagree), or have an even number of response options, such that a middle category, (e.g., Neither Agree nor Disagree), is not included as an option. The inclusion of the middle category on the NCTWCS could not be collapsed, and is therefore presented as a limitation because the category parameter estimates associated with the middle response option, especially for the unfolding model, are difficult to interpret *at the subjective level* (i.e., neither agree nor disagree from above, neither agree nor disagree from below).

Ideally, in any measurement situation, test construction should be guided, a priori, by a clear definition of the construct, the intended purpose of the survey, scoring methods, and intended use of test scores. Model selection for scaling/scoring requires the consideration of the assumptions that underlie the data, the assumptions that underlie the response process, the item response categories (i.e., yes/no or a gradation of categories as those in the current investigation), the relationship between the categories (i.e., ordered, unordered, interval), the intended uses of the scores, and to a lesser extent, sample size. Data to which an unfolding IRT model is appropriately applied, would be generated from a test built within the Thurstone framework where items, measure the entire spectrum of a single, clearly defined, non-cognitive construct. Items written in the Thurstone framework, naturally, do not include ambivalent (i.e., neither agree nor disagree, or no opinion) response options. Further, if it is assumed, a priori, that an ideal point response process would guide respondents when answering items, then application of a model that holds the same assumptions (i.e., unfolding IRT models) would be ideal. Items that measure attitudes, opinions, and perspectives, for which a response can be provided for two reasons (i.e., agree from below or agree from above) would necessitate the use of unfolding models. When observed responses to all items are plausibly provided for a single reason (i.e., there is no ambiguity surrounding responses) then application of an unfolding model would not be superior in measuring the construct over other scaling/scoring methods. Rather, unfolding models would likely not be selected as a scoring method because it would violate the principal of parsimony, and would artificially appear to fit the data better than other scoring methods simply due to over-parameterization.

## Summary and Interpretation of Results

The first research question related to examining the location of the items on the underlying latent traits (i.e., teacher perception of Empowerment and Leadership) across three IRT scaling methods (partial credit model, generalized partial credit model, and the generalized graded unfolding model). Item location can yield information about the comprehensiveness of latent trait measurement. It has also been argued by Chernyshenko et al. (2007), that unfolding IRT item parameters are directly associated with item content, whereas the item location parameters generated from cumulative IRT models are not. Regardless of the model used, item parameter estimates have implications for scale development.

In addressing the first research question, it was presumed that if the items on the attitudinal measure were constructed using a Likert methodology and modeled using an unfolding model assuming an ideal point response process when individuals respond to items, that items would generally be located at the ends of the latent trait continuum. Neutral items would not appear on the NCTWCS; items meeting the Likert criteria tend to be worded in more extreme terms (i.e., items that express both strong positive and strong negative sentiment with respect to the latent trait). It was hypothesized that if the items were not constructed using a Likert methodology, then item locations on the latent trait would be more similar to each other across all three scoring and scaling methods, than if a strict Likert methodology were used. Specifically, item locations would generally be more centrally located, or at least more dispersed across the attitude continuum as opposed to located towards the extreme values of the latent trait.

In the Empowerment analyses, there was some dispersion of the item locations generated from the PCM and GPCM models with a range of the averaged (across the 10 samples) locations of between -.936 to 1.823 and -1.034 to 2.390, respectively. Likewise, the locations of the items resulting from the GGUM (reversing the initial sign of the item locations during item parameter estimation) were slightly dispersed, with a range of the averaged item locations between 2.738 to 5.048. The GGUM location estimates were all relatively extreme and had very large associated standard errors for all 13 item location parameter estimates. In the Leadership analyses, the item locations were more tightly clustered in one region of the latent trait by all three IRT models, especially the GGUM.

Based on item location estimates, it is clear that only a portion of the attitude continuum is being measured by the 21 Leadership items. The 13 Empowerment items measure a slightly broader range of the latent trait. It is important it keep in mind that, especially for the 21 Leadership items, more than half of the over 65,000 respondents agreed or strongly agreed with each of the Leadership items. Again, as evidenced by the negative item locations (for the cumulative IRT models) combined with the descriptive information for these items, it was easy for the respondents to agree with the Leadership items, and most of them did. Generally, results from all three IRT models reveal that only a portion of the latent traits Leadership and Empowerment are being measured by these scales. Based on the rank order correlations, the two cumulative models ordered the Empowerment items identically, but both cumulative models produced highly dissimilar rank orderings of item locations when compared to the GGUM location estimates. The Leadership results, however, revealed that the item location estimates from all three IRT models were highly correlated. Both the PCM and the GPCM location estimates correlated highly and significantly with the GGUM estimates, meaning that Leadership items were ordered very similarly across the three IRT models. However, all IRT models estimated the items to be located in a narrow region of the trait. This point should be considered concurrently with the high rank order correlations.

A final word on the interpretation of the relative extremity of the item location estimates, especially those generated from the GGUM. To reiterate, the GGUM is a proximity based model, and the extremity of items is relative to the location of the attitude of the sample. It is possible that the Empowerment and Leadership items are moderately positive in content, but appear more extremely located if the respondents have more neutral attitudes relative to the items. In essence, homogeneity of the sample effects the extremity of item locations in proximity (i.e., unfolding) models.

The item discrimination parameters are useful to examine for the purpose of examining measurement precision and for future scale construction. Item discrimination parameters usually fall between values of 0.0 and 2.0 (Hambleton & Swaminathan, 1985). On average, most of the Empowerment items had very low discrimination values, meaning that they did not differentiate well among respondents of varying levels of the latent trait. The Leadership items yielded generally higher discrimination values on average. Item threshold/step parameters are useful because they demonstrate or provide a sense of how well respondents are using the response categories. Threshold parameters that have close estimates, like those in the Leadership analyses for the Disagree/Neither Agree Nor Disagree and the Neither Agree Nor Disagree/Agree, are evidence that these three categories were not used equally across respondents. The Empowerment items showed better category use than the Leadership items.

The second research question dealt with the person trait estimates resulting from the three IRT and the CFA models. The intended plan of analysis was to focus on the ends of the latent trait scale, as this is where cumulative and unfolding models are most discrepant. Analysis results revealed however that it was not completely necessary to emphasize so much on the estimates at the ends of the latent trait scale across models. Because the IRT models functioned very similarly for most of the items on both the Leadership and Empowerment scales, it was not necessary to focus so much on the extremes. This functioning can be seen in the scatterplots of the trait estimates (and to some extent the category probability plots as well). The scatterplots of the person trait estimates with the GGUM revealed that the PCM and GGUM were mostly consistent in their estimates with a linear relationship. The GGUM and GPCM estimates were almost perfectly linear and coincident. The GGUM and CFA person estimates were slightly nonlinear. None of the scatterplots displayed what would typically be seen if indeed the NCTWCS data were truly of the unfolding-type and measured with a cumulative-type model. If data were of the unfolding type, then the scatterplots of the person estimates would have revealed that traits generated from the cumulative-type models would be depressed, or show less extreme estimates, than the GGUM parameters. Specifically if the GGUM produced the "true" person estimate, then the expected scatterplot of the GGUM and CFA (or PCM, GPCM) estimates would have appeared to be "an elongated S-shaped function relating the two measures" (Roberts et a., 1999, p. 221). Further, the GGUM would have yielded more extreme person estimates than the CFA or cumulative IRT methods at the ends of the distribution if, and only if the true estimates were produced from the ideal point response process, and if and only if the entire continuum of the latent trait was measured (i.e., scale construction mimicked a Thurstone approach to development). .Consequently, there was little to examine and pay close attention to at the extremes of the latent trait distributions. When discrepancies did exist, they did so along the trait scale, not just at the ends. In reviewing the scatterplots and cross tabulation tables, the GGUM only yielded higher person trait estimates when compared to the PCM with the Leadership data. Specifically, as tabled in Chapter 4 for the first sample only, GGUM clustered 44 respondents into the 4th quintile, where those same people were categorized as falling into the 3rd, according to the PCM.

Examination of the probability function plots was the focus of the third research question. Any evidence of non-monotonicity of the probability functions from the cumulative IRT models would support the possibility that the ideal point response process was operational. Any monotonicity of the probability functions from the GGUM would indicate that either the cumulative and unfolding models are measuring the construct similarly, or that the items are extremely located relative to the respondents. It was hypothesized that, because the items on both scales seemed to contain relatively neutral content, the probability plots generated from the unfolding model would be single-peaked. Because the items on both scales did not appear to be extremely worded, or require an extreme (in either direction) attitude towards the latent trait for endorsement, it was hypothesized that the ICCs associated with the neutral items would exhibit folding, and this single-peaked nature would be marked in the GGUM analyses. This was not the case, however. The category probability plots from all three IRT models were characteristic of cumulative IRT models. The only items that showed unfolding characteristics, based on the GGUM analyses, were the first four Empowerment items, and items 7, 11 through 17, and 20 in the Leadership analyses at the upper end of the latent trait. The plots for these items, however, show only minimal unfolding properties for the categories of Strongly Agree and Agree. Across both scales, there was only a single item that exhibited unfolding properties across all categories, and that was the last item on the Leadership scale. That item reads: “Overall, the school leadership in my school is effective.” Further investigation as to why only this item may have generated data that needed “unfolding” is warranted. On average, across the 10 samples, this item was the second least discriminating, compared to the other 21 items. Of the 65,031 total respondents, 64,845 (99.7%) answered this question where 9.1% strongly disagreed, 13.0%

disagreed, 14.9% neither agreed nor disagreed, 44.4% agreed, and 18.3% strongly agreed. Although the plots from the first sample are tabled in Chapter 4, category plots from the other nine samples actually reveal a less marked display for item 21, where only the Strongly Agree and Agree response options exhibit unfolding characteristics. The category probability plots for Leadership item 21 from the nine samples are presented in Appendix B.

It was also hypothesized that the probability plots generated from cumulative and unfolding models would be the same for extremely worded items or would result from a homogeneous sample. This claim was supported, especially in the Leadership analyses, in that across the GPCM, PCM, and GGUM analyses, the Leadership items were moderately to extremely located on the latent trait, because most people agreed or strongly agreed with the items. Also the large standard errors associated with the GGUM location estimates imply a homogeneous sample relative to the item location. Therefore, the cumulative and unfolding models generated probability plots that were very similar.

The fact that most of the category probability plots were difficult to distinguish between the cumulative and unfolding models does not necessarily mean that the ideal point process was not in operation or responsible for the observed data. However, the findings from the first two research questions, coupled with the category probability plots suggests that the two types of IRT models functioned similarly enough not to warrant close examination of the extreme ends of the probability functions. Such attention would have been warranted if discrepancies were found between the cumulative and unfolding models, however, this was not the case for the Empowerment and Leadership data.

The ICCs generated from the GGUM analyses were monotonic for the last nine Empowerment items and for all Leadership items, except item 21. As expected, the non-monotonicity for the four Empowerment items and one Leadership item was very slight at the upper end of the latent trait.  According to Roberts et al. (1999) when referencing unfolding IRT models, "the degree of monotonicity inherent in the ICCs will be highly dependent on the relative locations of persons and items on the attitude continuum, and the range of person locations can obviously change from sample to sample" (p. 231). Further Roberts et al. (1998) demonstrated that items with truly inherent non-monotonic characteristics exhibit monotonic features when the sample of respondents' attitudes is restricted. Given the Leadership analyses (and to a large extent the Empowerment analyses) reported in Chapter 4, most people agreed with the items and the category probability plots across the three IRT models were similar with the exception of a few items. This demonstrates that the sample was very homogeneous. Additionally, the items on the Leadership scale were estimated to be located in narrow region of the latent trait. These results indicate that attitudes of the sample members were homogeneous and that the latent trait is not estimated well. Neither items nor people were disbursed across the latent trait, so even if an ideal point response process were responsible for the observed data, the homogeneity of the sample would preclude overt evidence of unfolding properties. The results presented in Chapter 4 support the notion that "a restricted sample range may mask the nonmonotonic response characteristics of a given item so that its characteristic curve appears to be monotonically related to attitude" (Roberts et al., 1998, p. 1).

The fourth research question pertained to two related issues: those of model assumptions and model fit. The assumption of unidimensionality within the context of cumulative IRT models was tested on the entire sample with a confirmatory factor analysis with one factor. The results and fit statistics can be interpreted as relatively poor fit with higher than acceptable values for measures like the RMSEA, RMR, SRMR, and low values for the GFI and NFI for the Empowerment data. The single factor structure imposed on the Leadership data yielded statistics that showed better fit than the Empowerment data with RMR, SRMR, and NFI values hovering at their respective cut-points. However, the RMSEA was higher than acceptable and the GFI was lower than acceptable. Determination of the model fit (i.e., a single factor model including 13 Empowerment or 21 Leadership items), and thus unidimensionality, and to some extent item independence, included the computation of the root mean square residuals for each item. Smaller values represent better fit. Based on the possible comparisons, the GPCM fit equally as well at the item level as the PCM on the Empowerment data, but showed much better fit of the Leadership data. Overall, unidimensionality was likely violated in the Empowerment analyses, but appeared to be met with the Leadership data according to most of the fit indices.

Testing the unidimensionality assumption within the context of unfolding models requires different methods than those used for cumulative IRT models. In the present study, the principal components analyses with two components did not fit well for either data set. The structure of the data was examined by plotting the factor pattern coefficients resulting from the PCA. The plots did not reveal that the data “unfold” because they did not form a semi-circular pattern. For both the Leadership and Empowerment data, the pattern coefficients associated with items were generally clustered in the two quadrants and did not form a fan-like configuration that would indicate that two linear principal components explain the pattern of data, and that the data unfold. A semi-circular plot of the pattern coefficients is evidence of unidimensionality within the context of unfolding models. Examination of the eigenvalues from the PCA also did not indicate that two components explained the data well. However, all final communality estimates were greater than 3. Considering all results, the unidimensionality assumption required for confident use of the GGUM also did not appear to be met for either data set.

Item and scale level chi-square distributed statistics are commonly used and reported in the educational and psychological measurement literature to test model-data fit. Such statistics were computed at the item scale level for both the Empowerment and Leadership measures, across all three IRT models. Two additional measures of model-data fit were calculated that do not demand absolute interpretations and are either insensitive to or control for sample size. The first was through the calculation of the root mean square residuals for each item and the second was with the calculation of the Akaike Information Criterion (AIC) and the Bayes Information Criterion (BIC). Both are measures of relative fit and smaller values are interpreted as better fitting. According to the chi-square statistics, neither the PCM, the GPCM, nor the GGUM fit the Leadership or Empowerment data well. Using the root mean square residuals at the item level, the GPCM appeared to fit the Leadership data better than the PCM. For the Empowerment data, however, the GPCM fit the data equally relative to the PCM. According to the AIC and BIC criteria, the GGUM fit both data sets relatively better than the GPCM, which fit both data sets relatively better than the PCM, though not by much. The question still remains, however, what “how much” means for these statistics and for differences between them.

## Empowerment

Information on how the NCTWCS was developed in terms of the decision making rules that governed how items eventually were determined to appear on the survey indicates that the NCTWCS 2006 responses were scored using confirmatory factor analytic methods to create domain scores for each sub-scale, including Empowerment and Leadership (Center for Teaching Quality, n.d.). If the sub-scales were also built using such approaches, then, presumably, the CFA analyses would have yielded better fit statistics, especially for the Empowerment scale. Further, the cumulative IRT assumption of unidimensionality presumably would have been met, again, especially for the Empowerment data.

It seems reasonable to conclude that results from the Empowerment analyses are questionable. This is because the single factor model using the CFA scaling approach did not fit; the assumption of unidimensionality was not met well (i.e., lack of good fit of a single factor structure) for the two cumulative IRT models; and because the assumption of unidimensionality within the context of IRT unfolding models also was not met. The potential consequences of violations of IRT model assumptions include biased item parameters and inaccurate person parameter estimates. This obviously has negative consequences for the reliability of estimates and validity of decisions based on those estimates. Further, violations prohibit or interfere with the comparison of individuals and individual differences on the latent trait, although this was not an intended or actual use of these data.

An interesting finding did emerge, however, with the Empowerment analyses. The first four items, and to a lesser extent, the fifth, exhibited some unfolding properties for the strongly agree and agree response options. Respondents were instructed to rate their level of agreement with the first five items on the Empowerment scale. For the last eight Empowerment items, teachers were instructed to indicate how large a role teachers have, and then were presented with a list of tasks. The response options associated with those eight items are: No role at all, small role, moderate role, large role, and the primary role. These options measure frequency more than level of satisfaction with or attitudes about the construct. Therefore, these eight items probably require a dominance response process, which is more than likely what respondents used to answer these items. Measures of frequency are most appropriately measured with cumulative IRT models (or at least models that assume a dominance process). In this case, then, because unfolding models, by design, can adequately measure cumulative-type data (i.e., fit monotonically increasing item response functions) as well as non-monotonic item response functions, the unfolding IRT model could theoretically be used for the Empowerment data. However, considering all of the results and that the presence of unfolding properties was not marked in the first five items, a model that presumes a dominance response process may be appropriate for these 13 Empowerment items.

Because questions about attitude and questions about frequency arguably measure different constructs, one consideration for future versions of the NCTWCS may be to carefully define the construct of interest and perhaps use separate scales for separate constructs. The “mixed” data resulting from the two types of questions (i.e., attitude, frequency) is likely the reason for the poor model-fit. Further, a two-dimensional structure was probably not evident either because a total of 13 items comprised the entire Empowerment scale, where five measured attitude and eight measured frequency. Although there is no absolute criterion for minimum (or maximum) number of items, in research investigations using real and simulated data, the number of items measuring a single dimension (i.e., construct) is rarely less than 10. A small number of items measuring a single construct yield low reliability. Another suggestion, then, would be to include additional items on the Empowerment survey, especially on the sub-scale that measures attitudes.

Method of scoring is necessarily an important, and must be considered in conjunction with the construct, purpose of the survey, and intended uses of survey scores.

## Leadership

A single factor confirmatory factor analytic model fit the Leadership data reasonably well, whereas the principal components analysis with two factors did not. However, the cumulative and unfolding models functioned quite similarly in terms of the rank ordering of Leadership item location and discrimination parameters, and in the rank order of person parameter estimates. The correspondence between the GGUM and GPCM for the theta estimates was also quite high. Chi-square distributed statistics for all models at the scale level would suggest that none of the models fit well, but according to the information theory-based statistics, the GGUM fit better than the GPCM. It would be necessary to explain or try to understand why the fit at the item and scale level was poor for all IRT models. Possibilities include the inappropriateness of the models for the data; the construct of "teachers perceptions of school leadership," is not measured well by the collection of items that comprise this scale; or that the sample size of 2,000 increased the degrees of freedom to the point of model rejection. The chi-square statistics should not be ignored, although decisions should not rest solely on these statistics. Although the CFA model cannot be directly compared with IRT models, given all of the findings, IRT models could reasonably be considered for at least the Leadership data. Reasons for preferring IRT models over factor analytic models include the fact that the former estimate item characteristics like discrimination, location, and category thresholds. Further, although not discussed in this study,  IRT models yield an index called information. This index is closely related to the discrimination parameters and provides an indication where on the latent trait distribution the item best measures, conditioned on theta. Information at the item level can be summed for a total test information index. So, although a determination cannot be made from this investigation which model is correct (CFA, PCM, GPCM, GGUM), at the very least for reliable and efficient scale construction, IRT models would be favored. The PCM, or any Rasch IRT model may not be preferred for this Leadership data because of the difficulty in fitting the PCM, and the associated low *a* parameters. Because of the rather homogeneous sample of attitudes and general clustering of item locations in a small region of the latent trait, the GGUM parameter estimates were generally extreme and were associated with high standard errors. As a result, for these 21 Leadership items, the GPCM might be preferred over the PCM and the GGUM.

## Implications

This investigation was conducted after NCTWC scale construction and data collection. This approach is not entirely ideal, however. Researchers such as Chernyskenko et al., (2007), Meijer and Baneke (2004), Stark et al., (2006) have applied unfolding IRT models to scales constructed using Likert-type approaches as a way to both examine cumulative IRT model assumptions and to investigate the applicability and appropriateness of unfolding IRT models. Argument supporting such an approach is that unfolding IRT models are flexible and versatile enough to model constructs that were measured using instruments designed using cumulative methods. If the NCTWCS was constructed using a Likert-type approach then presumably no items should exhibit unfolding properties. However, some unfolding attributes were evident in both Leadership and Empowerment analyses, specifically for the Strongly Agree and Agree response options and for all five response options in the last Leadership item.

One question that this investigation attempted to answer is whether or not unfolding IRT models would be an appropriate alternative to CFA or cumulative IRT models for the current 13 Empowerment and 21 Leadership items. It appears that the data are not of the unfolding type based on the results when examining unfolding IRT assumptions and the structure of the Leadership and Empowerment data. Model assumptions were also violated using the Empowerment data for the CFA and cumulative IRT models. As a general rule when considering the application of any mathematical model, if assumptions are violated, analyses should be conducted to determine why the violations occurred and appropriate actions taken.

The results in this investigation can contribute to future versions of the NCTWCS. There were no items on either scale that contained extreme content in either direction (e.g., “I would not change a single aspect about the leadership in my school” or “The poor leadership in my school contributes to difficulties in retaining teachers”). The items were not necessarily ambivalent either (e.g., “Sometimes I agree with the decisions and processes imposed by the leadership in my school and sometimes I do not”). Across both scales, items did not appear to tap the entire spectrum of a highly positive attitude towards current school leadership (empowerment) or include items that would require disenchantment with leadership (empowerment) and low levels of attitude towards the construct. At the item level all IRT models generally estimated the Leadership items to be located in one general region of the scale with slight dispersion of the Empowerment items. The NCTWCS was developed to understand how the population of teachers in North Carolina perceive their work environment, and to gain insight as to what is not working well in that environment. The intention of the Office of the Governor was using the data to make the necessary changes to school leadership to improve working conditions. Improvement of working conditions was assumed to be directly related to increased student learning and achievement, and by improving teacher working conditions, improved student learning and achievement would follow.

According the model results, and even the descriptive information (i.e., percentage of people responding to each category of each item) it appears that the Leadership items are easy to endorse, that most people agreed or strongly agreed with them, and that the item locations were estimated to be clustered closely together. Lack of good category response option usage is an undesired consequence as a data set that consists of strongly agree and agree responses results in a lack of variability across respondents making small differences in person trait estimates difficult to find. Further, and perhaps a more unfavorable consequence of a lack of variability is the suppression of information about respondents standing on the latent trait. Such items are not useful for reliably measuring an individual’s standing on the latent trait, and for finding individual differences among respondents on the latent trait..

For future revisions of the NCTWCS it may be more efficient to add items that do indeed measure characteristics of school leadership (empowerment) that may not be so positive, and perhaps remove some of the current items, as there is so much overlap among them. The items currently comprising these two scales still do not measure dissatisfaction with school leadership (Empowerment), an apparent contributing factor of teacher turnover. The results give the Office of the Governor little to “improve.”

The implications of this investigation stretch beyond the NCTWCS data to include implications for survey development (for the assessment of non-cognitive constructs), and for methodologies typically used to assess the appropriateness and functioning of unfolding IRT models. As for survey development, some of the initial considerations in the development process should include a clear purpose of the survey and a clear definition of the construct to be measured. The processes that people use to respond to items must also be taken into account during the item writing process. Further, consideration of scoring methods should be concurrent with the initial steps of the test development process. Because the NCTWCS was presumably constructed using a Likert methodology, because the items did not read as though they could be endorsed for two reasons, and because items did not measure a range of the underlying traits, a cumulative approach likely would be the most appropriate and consistent scaling method. Use of a survey with known psychometric properties and test development strategies that are aligned with the scaling method (i.e., unfolding IRT model) would have facilitated appropriate model selection.

The consideration of scoring methods during the test development process is related to the implications for methodologies for research the surrounds the functioning and applicability of unfolding IRT models. In the present study, scoring models were applied post-hoc. However, this is not typically the case; rather, the scoring model is selected prior to test construction. A more sound approach for assessing the functioning and appropriateness of unfolding IRT models would be to construct a scale within the Thurstone framework, (given that this method for scale construction is aligned with the purpose of the survey), then apply the appropriate scoring model.

## Suggestions for Future Research

The report detailing the findings of the 2006 NCTWCS data (Hirsch, Emerick, Church, & Fuller, 2006) opens with assertions regarding the importance and influence that teachers have on students. The issue of teacher turnover is immediately noted and described as a problem, and negative consequences of turnover are explained. The intent of the NCTWCS was to get a sense of how teachers feel about their work environment, all for the purpose of helping teachers help students learn. The effect would be increased student achievement. Based on publicly available documentation regarding the survey, the links between teachers' perceptions of working conditions and improved student achievement are such that, if teachers are happy and comfortable in their working environment, they are apt to stay in that environment. If teachers are happy in their jobs and the climate of the school, they are likely to be satisfied. A satisfied teacher is assumed to be motivated, thus leading to more effective teaching as evidenced by the teacher having a positive impact on student achievement. Compared to teachers who leave a school or the profession, teachers who stay have a positive impact on the classroom environment, reduce disruption, and retaining teachers can yield economic benefits for a school district.

If the purpose of a scale or survey is to elicit attitudes towards some construct to get a better sense about attitudes of a sample, the scale needs different items that measure the full spectrum of the latent trait. This would more than likely increase the variability among respondents allowing as much information as possible to be gained from each item. Items that measure a range of attitudes certainly would increase the efficiency of the survey because little is achieved if most items are only measuring one region of the full latent trait spectrum. Several items that measure the same exact point on the latent trait continuum are not entirely useful or informative; they are not a good use of time for those writing the items, responding to the items, and for those scoring the items. As it stands, most of the Leadership items were easy to endorse. This yields little variation among respondents and essentially inhibits any conclusions about which aspects of Leadership can be improved to improve teachers’ perceptions of the leadership in their school, reduce teacher turnover, and improve student learning and achievement.

One component that seemingly was not considered in the original data collection and use nor in the current investigation is that of nested data. Teachers were surveyed, although decisions were made at the school level. The results of the 2006 NCTWCS administration were used to name and award 10 schools across the state. The North Carolina Professional Teaching Standards Commission and the Governor’s Teachers Advisory Committee designated these as 2006 Real D.E.A.L (Dedicated Administrators, Educators, and Learners) schools which were determined to function as exemplars for best practice for other districts in terms of teacher working conditions. Therefore, aside from modifying the NCTWCS to include items that measure more of the latent trait and eliminate redundant items, methodologically, it makes sense to either examine the consequences of ignoring the nesting or to proceed with future survey administrations and analyses that accommodate the nested structure of the data.

The hypothesis described previously regarding how satisfied employees who are happy and comfortable in their occupation and place of employment are likely to be more productive and efficient is probably true of most employees, regardless of job title, job requirements, or field. Although the data used here were from the NCTWCS, any employer seeking to make improvements in the work environment with the intent of increased employee productivity and efficiency would be best served by doing so using a scale that is comprehensive and efficient in its measurement. The purpose of the scale and the intended use of the scores should necessarily guide the scale development and scoring processes. This point could not be investigated in the current study because the NCTWCS was not constructed using a Thurstone-type approach.  However, if the purpose of the assessment is to ascertain individual level information about individual attitudes or preferences, it seems feasible--theoretically and practically--that unfolding IRT models could be used.

Aside from making decisions about educational policy or employee satisfaction, another potential practical application of unfolding models for preference or attitude data, is for making decisions about or informing occupational decisions with the use of practice analyses. A practice analysis (sometimes referred to as job analysis) is the examination of knowledge, skills, and abilities required of a particular occupation in a particular industry. A practice analysis begins with a survey of the importance and relevance of both current requirements of an occupation and perhaps additional and novel capabilities introduced to the profession. Not only is the importance of each skill, ability, or task measured by way of rating scales, but so are other characteristics such as frequency of each task/skill. Practice analyses are often conducted as a way to examine such things as content validity of educational and training programs for particular occupations. Because non-cognitive constructs like importance and critical nature of tasks are measured together with cognitive aspects like frequency of those skills/tasks, it seems that application of a measurement model than can accommodate these types of measurement situations would be most appropriate. To date, parametric and non-parametric unfolding IRT models have been shown to reliably accommodate response processes that produce observed data when measuring both types of constructs with dichotomous and/or polytomous items.

For unfolding IRT models, more research is useful and necessary in an applied sense (i.e., measuring different constructs, populations, instruments), the paucity of research that focuses on the technical aspects of the structure of unfolding data is cause for added scrutiny, investigation, testing, and understanding, if unfolding IRT models are to be incorporated into applied research situations. Explanations have been provided by Davison (1977), van Schuur and Kiers (1994) and Mauran and Rossi (2001) as to how and why application of linear factor analytic methods to unidimensional, unfolding-type data reveals a two-factor structure. Davison (1977) further explained and showed what the inter-item correlation and partial correlation matrices look like for unidimensional data that fit an unfolding model. Data structure is necessary to consider and understand as it is directly related to the assumption of dimensionality. Future research should address the structure of unidimensional data, and the relationship between item responses within an ideal-point framework as consensus does not exist among researchers regarding the issue of dimensionality structure of unfolding-type data and methods for assessing (uni)dimensionality of unfolding-type data. For example, according to Roberts et al. (1996), Roberts et al. (2000), and Nandakumar et al. (2002) if a two-factor structure results from the application of a principal components analysis, and if item level communalities are greater than .3 (from the first 2 components), then the data are considered unidimensional, of the unfolding type. On the other hand, the methods (and criteria) used to assess dimensionality within the cumulative framework by Chernyshenko et al. (2007) were also used for assessing and determining dimensionality within the unfolding framework. They used linear factor analytic methods to assess dimensionality and if a single factor model fit the data, then unidimensionality was assumed from both a cumulative and ideal-point perspective. Their reasoning rests on the fact that they could not find any simulation studies that tested the accuracy of the former approach and criteria (i.e., two factor structure resulting from PCA; item level communalities greater than .3). A greater understanding of unfolding-type data structure (inter-item correlation and covariance matrices), and a resolution to this methodological issue would be facilitated with simulation studies. Such efforts could further inform the development of statistical tests for determining the dimensionality structure of unfolding-type data. Once a better understanding of unidimensional unfolding data is gained, endeavors should be extended to the multidimensional measurement situation. Currently, little is known about data that would be appropriate for multidimensional, unfolding, IRT models. Further, far fewer multidimensional models exist for measuring multidimensional non-cognitive data, where multidimensional, unfolding, IRT models are still to be developed.

## Conclusion

Similar to what some researchers have found in examining characteristics of items on attitudinal surveys, the results from this investigation show that some items on both attitude scales possessed unfolding-type properties. On the Empowerment scale, only those items that measured attitude displayed such properties, while those that measured frequency showed cumulative-type properties. On the Leadership scale, close to one-third of the items displayed some unfolding characteristics. These findings provide some indication that the ideal point response process may be, at least in part, responsible for the observed data. This finding is not necessarily a reason to suggest the immediate implementation of unfolding IRT models in scoring future administrations of the NCTWCS, but this does have implications for survey design and scoring. In the presumed current scoring method, a sum score is used. This implies a cumulative framework where a higher score implies more of the construct. If higher is not more, for some items, then a simple summation is inaccurate and would likely inflate survey scores.

Due to the clustering of items on both scales, and considering the purpose of the survey and the policy decisions that are made using the survey scores, it may be useful to remove redundant items, add more items that measure different attitudes towards working conditions, use separate scales for different constructs, and carefully consider the both the survey development and scoring procedures. Because of the frequency with which surveys are administered eliciting individuals' opinions for the purpose of making various decisions (hiring, production/sales/marketing, assessing symptoms of psychiatric illnesses, evaluating programs and services), considering that the surveys contain Likert-type items and employ a Likert-type procedure for scoring, recommendations for the simultaneous consideration of the construct, scale development, and scoring procedures are not unique to the NCTWCS, but to all related surveys with similar intentions.

APPENDIX A

# 2006 North Carolina Teacher Working Conditions

### Thank you in advance for your time and willingness to share your views on working conditions in your school.

Research has demonstrated that teacher working conditions are critical to increasing student achievement and retaining teachers. North Carolina policymakers and education stakeholders have expressed great interest in using your collective responses on this survey to help improve working conditions in schools and districts across the state.

### Please know that your anonymity is guaranteed.

No one in your school, the district or state will be able to view individual surveys, and reports on the results will not include data that could identify individuals. You are being asked demographic information to learn whether teachers from different backgrounds and different characteristics look at working conditions differently.

### Access Code

You have been assigned an anonymous access code to ensure that we can identify the school in which you work and to ensure the survey is taken only once by each respondent. The code can only be used to identify a school, and not an individual. The effectiveness of the survey is dependent upon your honest completion.

**Please indicate your position:**

Teacher (including intervention specialist, vocational, literacy specialist, special education, etc.)
Principal
Assistant Principal
Other Education Professional (school counselor, school psychologist, social worker, library media specialist, etc.)

# Time

Please rate how strongly you agree or disagree with the following statements about the use of time in **your school**.

**Please indicate your level of agreement with the following statements.**

a. **Teachers*** have reasonable class sizes, affording them time to meet the educational needs of all students.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

b. Teachers have time available to collaborate with their colleagues.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

c. Teachers are protected from duties that interfere with their essential role of educating students.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

d. School leadership tries to minimize the amount of routine administrative paperwork required of teachers.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

e. The **non-instructional time*** provided for teachers in my school is sufficient.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

**"Teachers" means a majority of teachers in your school.*

**"Non-instructional time" refers to any structured time during the work day to work individually or collaboratively on instructional issues.*

**In an average week of teaching, how many hours do you have for non-instructional time during the regular school day?**

None
Less than 3 hours
More than 3 hours but less than or equal to 5 hours
More than 5 hours but less than or equal to 10 hours
More than 10 hours

**In an average week of teaching, how much non-instructional time do teachers have available?**

None
Less than 3 hours
More than 3 hours but less than or equal to 5 hours
More than 5 hours but less than or equal to 10 hours
More than 10 hours

**Of those hours, how many are available for individual planning?**

None
Less than 3 hours
More than 3 hours but less than or equal to 5 hours
More than 5 hours but less than or equal to 10 hours
More than 10 hours

**And how many hours are available for structured collaborative planning?**

None
Less than 3 hours
More than 3 hours but less than or equal to 5 hours
More than 5 hours but less than or equal to 10 hours
More than 10 hours

**In an average week of teaching, how many hours do you spend on school related activities outside the regular school work day (before or after school, and/or on the weekend)?**
None
Less than 3 hours
More than 3 hours but less than or equal to 5 hours
More than 5 hours but less than or equal to 10 hours
More than 10 hours

**In an average week of teaching, how many hours do teachers spend on school-related activities outside of the regular school work day?**

None
Less than 3 hours
More than 3 hours but less than or equal to 5 hours
More than 5 hours but less than or equal to 10 hours
More than 10 hours

# Facilities and Resources

Please rate how strongly you agree or disagree with the following statements about your school facilities and resources.

a. Teachers have sufficient access to appropriate **instructional materials*** and resources.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

b. Teachers have sufficient access to instructional technology, including computers, printers, software, and internet access.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

c. Teachers have sufficient access to communications technology, including phones, faxes, email, and network drives.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

d. Teachers have sufficient access to office equipment and supplies such as copy machines, paper, pens, etc.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

e. The reliability and speed of Internet connections in this school are sufficient to support instructional practices.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

f. Teachers have adequate professional space to work productively.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

g. Teachers and staff work in a school environment that is clean and well maintained.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

h. Teachers and staff work in a school environment that is safe.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

**Instructional materials include items such as textbooks, curriculum materials, content references, etc.*

## Teacher Empowerment

Please rate how strongly you agree or disagree with the following statements about teacher empowerment **in your school**.

**Please rate your level of agreement with the following statements.**
a. Teachers are centrally involved in decision making about educational issues.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

b. Teachers are trusted to make sound professional decisions about instruction.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

c. The faculty has an effective process for making group decisions and solving problems.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

d. In this school we take steps to solve problems.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

e. Opportunities for advancement within the teaching profession (other than administration) are available to me.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

**Please indicate how large a role teachers at your school have in each of the following areas:**

a. Selecting instructional materials and resources.
**No role at all Small role Moderate role Large role The primary role**

b. Devising teaching techniques.
**No role at all Small role Moderate role Large role The primary role**

c. Setting grading and student assessment practices.
**No role at all Small role Moderate role Large role The primary role**

d. Determining the content of in-service professional development programs.
**No role at all Small role Moderate role Large role The primary role**

e. Hiring new teachers.
**No role at all Small role Moderate role Large role The primary role**

f. Establishing and implementing policies about student discipline.
**No role at all Small role Moderate role Large role The primary role**

g. Deciding how the school budget will be spent.
**No role at all Small role Moderate role Large role The primary role**

h. School improvement planning.
**No role at all Small role Moderate role Large role The primary role**

**Members of the school improvement team are elected.**
Yes
No
Don't know

# Leadership

Please rate how strongly you agree or disagree with the following statements about leadership **in your school**.

**Please rate your level of agreement with the following statements.**

a. There is an atmosphere of trust and mutual respect within the school.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

b. The faculty are committed to helping every student learn.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

c. The school leadership communicates clear expectations to students and parents.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

d. The school leadership shields teachers from disruptions, allowing teachers to focus on educating students.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

e. The school leadership consistently enforces rules for student conduct.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

f. The school leadership support teachers' efforts to maintain discipline in the classroom.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

g. Opportunities are available for members of the community to actively contribute to this school's success.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

h. The school leadership consistently supports teachers.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

i. The school improvement team provides effective leadership at this school.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

j. The faculty and staff have a shared vision.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

k. Teachers are held to high professional standards for delivering instruction.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

l. Teacher performance evaluations are handled in an appropriate manner.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

m. The procedures for teacher performance evaluations are consistent.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

n. Teachers receive feedback that can help them improve teaching.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

**The school leadership makes a sustained effort to address teacher concerns about:**

a. facilities and resources
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

b. the use of time in my school
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

c. professional development
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

d. empowering teachers
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

e. leadership issues
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

f. new teacher support.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

Overall, the school leadership in my school is effective.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

**Which position best describes the person who most often provides instructional leadership at your school?**
principal or school head
assistant or vice principal
department chair or grade level chair
school-based curriculum specialist
director of curriculum and instruction or other central office based personnel
Other teachers
None of the above.

# Professional Development

Please rate how strongly you agree or disagree with the following statements about your own professional development and professional development in **your school**.

**Please indicate your level of agreement with the following statements.**

a. Sufficient funds and resources are available to allow teachers to take advantage of professional development activities.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

b. Teachers are provided opportunities to learn from one another.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

c. Adequate time is provided for professional development.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

d. Teachers have sufficient training to fully utilize instructional technology.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

e. Professional development provides teachers with the knowledge and skills most needed to teach effectively.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

**In which of the following areas, if any, do you believe teachers need additional support to effectively teach students?**

Special education (students with disabilities)
Special education (academically gifted students)
Limited English Proficiency (LEP)
Closing the achievement gap
Your content area
Methods of teaching
Student assessment
Classroom management techniques
Reading strategies

**In which of the following areas, if any, do you need additional support to effectively teach your students? Check all that apply.**

Special education (students with disabilities)
Special education (academically gifted students)
Limited English Proficiency (LEP)
Closing the achievement gap
Your content area
Methods of teaching
Student assessment
Classroom management techniques
Reading strategies

**In the past 2 years, have you had 10 hours or more of professional development in any of the following areas? Check all that apply.**

Special education (students with disabilities)
Special education (academically gifted students)
Limited English Proficiency (LEP)
Closing the achievement gap
Your content area
Methods of teaching
Student assessment
Classroom management techniques
Reading strategies

**Did the professional development you received in** special education for students with disabilities **provide you with strategies that you have incorporated into your instructional delivery methods?**
Yes
No

**Did the professional development you received in** special education for academically gifted students **provide you with strategies that you have incorporated into your instructional delivery methods?**
Yes
No

**Did the professional development you received in** LEP **provide you with strategies that you have incorporated into your instructional delivery methods?**
Yes
No

**Did the professional development you received in** closing the achievement gap **provide you with strategies that you have incorporated into your instructional delivery methods?**
Yes
No

**Did the professional development you received in** your content area **provide you with strategies that you have incorporated into your instructional delivery methods?**
Yes
No

**Did the professional development you received in** methods of teaching **provide you with strategies that you have incorporated into your instructional delivery methods?**
Yes
No

**Did the professional development you received in** student assessment **provide you with strategies that you have incorporated into your instructional delivery methods?**
Yes
No

**Did the professional development you received in** classroom management techniques **provide you with strategies that you have incorporated into your instructional delivery methods?**
Yes
No

**Did the professional development you received in** reading strategies **provide you with strategies that you have incorporated into your instructional delivery methods?**
Yes
No

**Were these strategies you learned in your professional development in** special education for students with disabilities **useful for your efforts to improve student achievement?**
Yes
No

**Were these strategies you learned in your professional development in** special education for academically gifted **useful for your efforts to improve student achievement?**
Yes
No

**Were these strategies you learned in your professional development in** LEP **useful for your efforts to improve student achievement?**
Yes
No

**Were these strategies you learned in your professional development in** closing the achievement gap **useful for your efforts to improve student achievement?**
Yes
No

**Were these strategies you learned in your professional development in** your content area **useful for your efforts to improve student achievement?**
Yes
No

**Were these strategies you learned in your professional development in** methods of teaching **useful for your efforts to improve student achievement?**
Yes
No

**Were these strategies you learned in your professional development in** student assessment **useful for your efforts to improve student achievement?**
Yes
No

**Were these strategies you learned in your professional development in** classroom management techniques **useful for your efforts to improve student achievement?**
Yes
No

**Were these strategies you learned in your professional development in** reading strategies **useful for your efforts to improve student achievement?**
Yes
No

**In the past two years, have you enrolled or participated in any of the following professional development activities?**

online learning opportunities
**Yes No**

local in-service program
**Yes No**

state-sponsored in-service program
**Yes No**

Was the Online learning opportunity required?
**Yes No**

The Online learning opportunities activities I participated in were effective.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

Was the local in-service program required?
**Yes**
**No**

The local in-service program activities I participated in were effective.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

**Was the state-sponsored in-service program required?**
Yes
No

The state-sponsored in-service program activities I participated in were effective.
**Strongly Disagree Disagree Neither Disagree Nor Agree Agree Strongly Agree**

Do you teach students who have an Individualized Education Plan or 504 Plan?
**Yes**
**No**

Do you teach students who are Limited English Proficient?
**Yes**
**No**

# Core Questions

**Which aspect of your work environment most affects your willingness to keep teaching at your school?**
Time during the work day
School facilities and resources
School leadership
Teacher empowerment
Professional Development

**Which aspect of your school's work environment most affects teachers' willingness to keep teaching at your school?**

Time during the work day
School facilities and resources
School leadership
Teacher empowerment
Professional Development

**Which aspect of working conditions is most important to you in promoting student learning?**

Time during the work day
School facilities and resources
School leadership
Teacher empowerment
Professional Development

**Overall, my school is a good place to teach and learn**
Strongly Disagree
Disagree
Neither Agree Nor Disagree
Agree
Strongly Agree

**At this school, we utilize results from the Teacher Working Conditions survey as a tool for Improvement.**

Strongly Disagree
Disagree
Neither Agree Nor Disagree
Agree
Strongly Agree

**Which BEST DESCRIBES your future intentions for your professional career?**

Continue teaching at my current school
Continue teaching at my current school until a better opportunity comes along.
Continue teaching but leave this school as soon as I can
Continue teaching but leave this district as soon as I can
Leave the profession all together

# Demographics

**Please tell us more about yourself. No demographic information that could be used to identify individual educators will be shared. All questions in this section are optional.**

**Please indicate your ethnicity.**
American Indian or Alaska Native
Asian or Pacific Islander
Black or African American
Hispanic
White
Mixed or multiple ethnicity
Some other race or ethnicity

**Please indicate your gender.**

Female
Male

**How did you train to become an educator?**
Bachelor's degree
Master's degree
Alternative route
**Highest degree attained**
Bachelor's
Master's
Doctorate
Other

**Are you certified by National Board for Professional Teaching Standards (NBPTS)?**
Yes
No

**How many years have you been employed as an educator?**
First Year
2 - 3 Years
4 - 6 Years
7 - 10 Years
11 - 20 Years
20+ Years

**How many years have you been employed in the school in which you are currently working?**
First Year
2 - 3 Years
4 - 6 Years
7 - 10 Years
11 - 20 Years
20+ Years

**Have you served as a mentor in North Carolina schools in the past five years?**
Yes
No

## Mentoring

**Have you been formally assigned a mentor in your first AND second year teaching in North Carolina?**
Yes
No

**Answer questions for a formal mentor assigned at the school where you now work. If you had multiple years of formal mentors, answer questions for your most recent mentor experience.**

**My mentor was effective in providing support in the following areas**

a. Instructional strategies
**Of no help at all   Has helped a little  Has helped some  Has helped a lot  Help was critical**

b. Curriculum and the subject content I teach
**Of no help at all   Has helped a little  Has helped some  Has helped a lot  Help was critical**

c. Classroom management/discipline strategies
**Of no help at all Has helped a little Has helped some Has helped a lot Help was critical**

d. School and/or district policies and procedures
**Of no help at all Has helped a little Has helped some Has helped a lot Help was critical**

e. Completing products or documentation required of new teachers
**Of no help at all Has helped a little Has helped some Has helped a lot Help was critical**

f. Completing other school or district paperwork
**Of no help at all Has helped a little Has helped some Has helped a lot Help was critical**

g. Social support and general encouragement
**Of no help at all Has helped a little Has helped some Has helped a lot Help was critical**

h. Other
**Of no help at all Has helped a little Has helped some Has helped a lot Help was critical**
**Please indicate whether each of the following were true for you and your mentor**
a. My mentor and I were in the same building(or school)
Yes
No

b. My mentor and I taught in the same content area
Yes
No

c. My mentor and I taught the same grade level
Yes
No

**On average, how often did you engage in each of the following activities with your mentor?**

a. Planning during the school day with my mentor

**Never Less than once/ month Once a month Several times / month Once a week Almost daily**

b. Being observed teaching by my mentor
**Never Less than once/ month Once a month Several times / month Once a week Almost daily**

c. Observing my mentor's teaching
**Never Less than once/ month Once a month Several times / month Once a week Almost daily**

d. Planning instruction with my mentor
**Never Less than once/ month Once a month Several times / month Once a week Almost daily**

e. Having discussions with my mentor about my teaching
**Never Less than once/ month Once a month Several times / month Once a week Almost daily**

f. Meeting with my mentor outside of the school day
**Never Less than once/ month Once a month Several times / month Once a week Almost daily**

**How important has your mentoring experience been in your decision to continue teaching at this school?**

Made no difference at all
Only slightly important
Somewhat important
Important
Very important

**If you have served as mentor in the past three years, please answer the following questions for YOUR MOST RECENT mentoring experience**

**Are you a full time mentor?**
Yes
No

**How many teachers did/do you mentor?**
1
2
3
4 - 6
7- 10
10 +

**On average, how often did/do you meet with your mentee(s)**
Never
Less than once per month
Once a month
Several times a month
Once a week
Almost daily

**Please indicate which best describes you and your mentee(s)**

a. My mentor and I were in the same building
**None of them   Some of them   All of them**

b. My mentor and I taught in the same content area
**None of them   Some of them   All of them**

c. My mentor and I taught the same grade level
**None of them   Some of them   All of them**

**On average, how often did you engage in each of the following activities with your mentee(s)?**

a. Planning during the school day with my mentee(s)
**Never  Less than once/ month  Once a month  Several times / month  Once a week  Almost daily**

b. Observing my mentee(s)' teaching
**Never  Less than once/ month  Once a month  Several times / month  Once a week  Almost daily**

c. Being observed by my mentee(s)
**Never  Less than once/ month  Once a month  Several times / month  Once a week  Almost daily**

d. Planning instruction with my mentee(s)
**Never  Less than once/ month  Once a month  Several times / month  Once a week  Almost daily**
e. Having discussions with my mentee(s) about teaching
**Never  Less than once/ month  Once a month  Several times / month  Once a week  Almost daily**

**Please indicate which of the following kinds of support, if any, you received as a formally assigned mentor. (Check all that apply).**

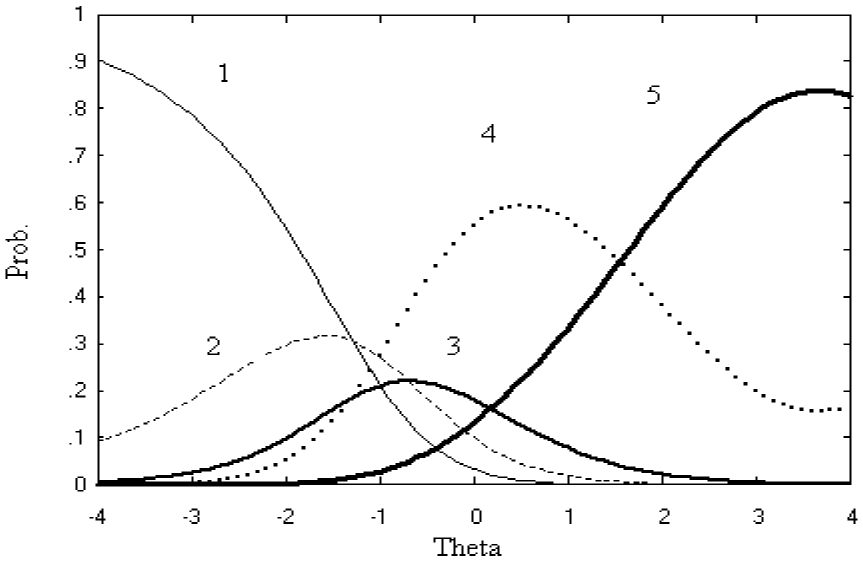
Release time to observe your mentee(s)
Release time to observe other mentors
Reduced teaching schedule
Reduced number of preparations
Common planning time with teachers you are mentoring
Specific training to serve as a mentor (e.g. seminars or classes)
Regular communication with principals, other administrator or department chair
Other

**Thank you for sharing your valuable time, thoughts and perspectives on this survey. We value the work you do to provide a quality education to the children of NC. Survey results will be available at http://www.northcarolinatwc.org by June 1, 2006.**

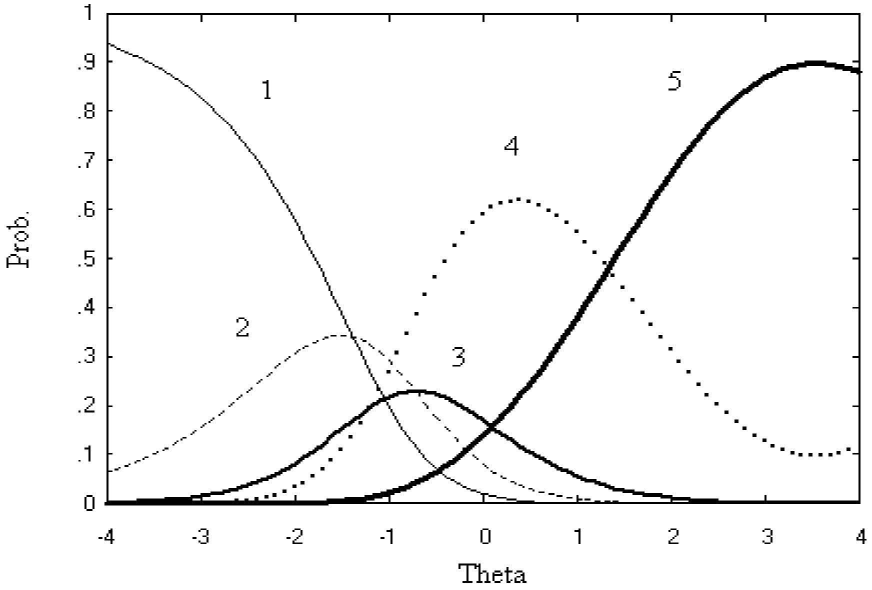
## APPENDIX B

## CATEGORY PROBABILTIY PLOTS FOR LEADERSHIP ITEM 21: SAMPLES 2-10

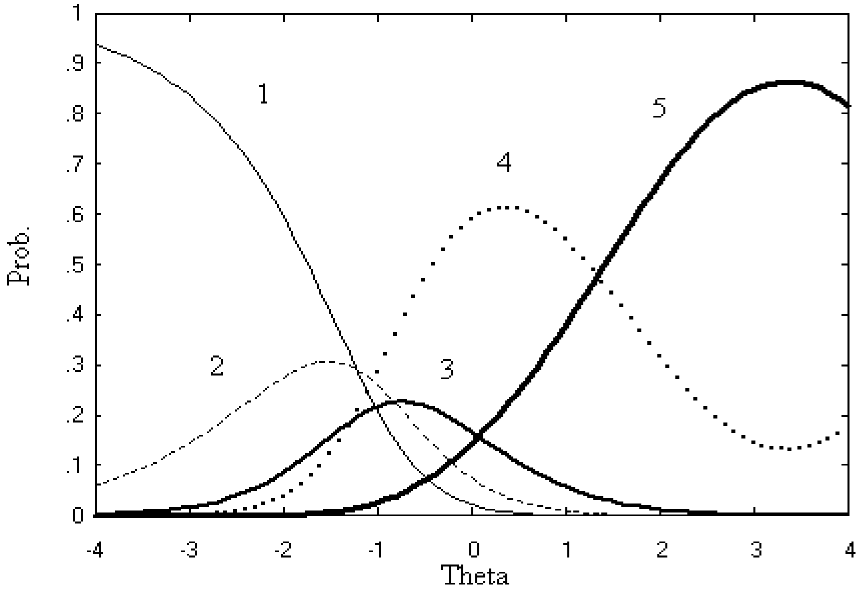
*Category Probability Plot for Item 21 with GGUM: Sample 2, Leadership Scale*



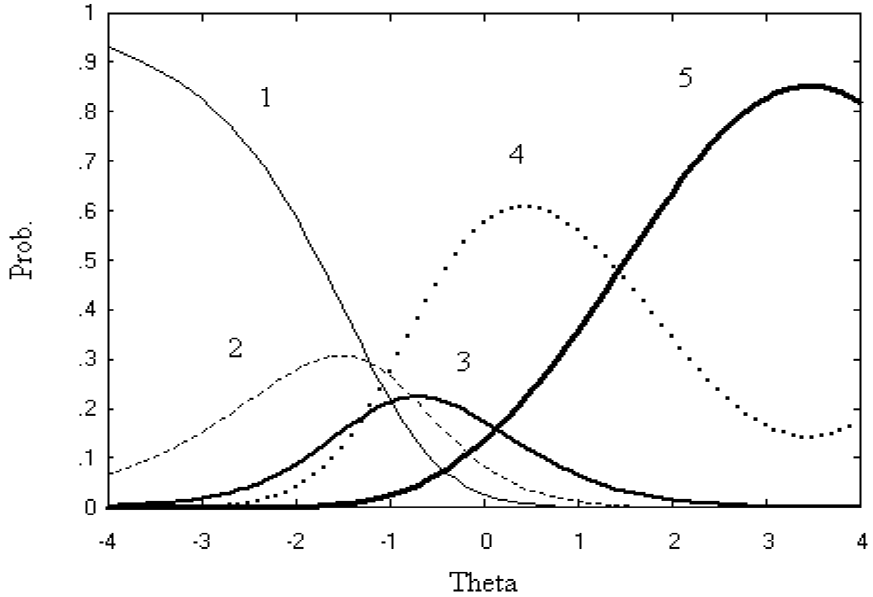
*Category Probability Plot for Item 21 with GGUM: Sample 3, Leadership Scale*



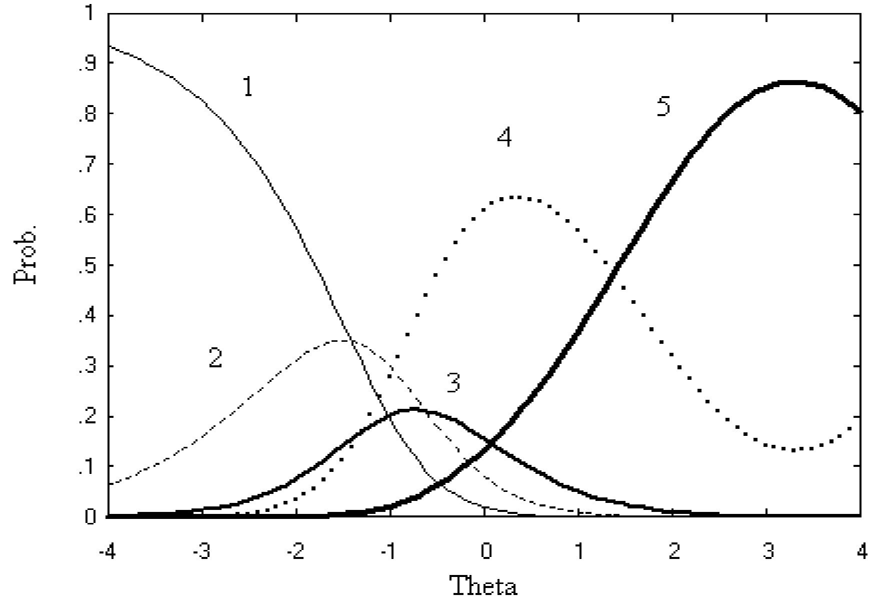
*Category Probability Plot for Item 21 with GGUM: Sample 4, Leadership Scale*



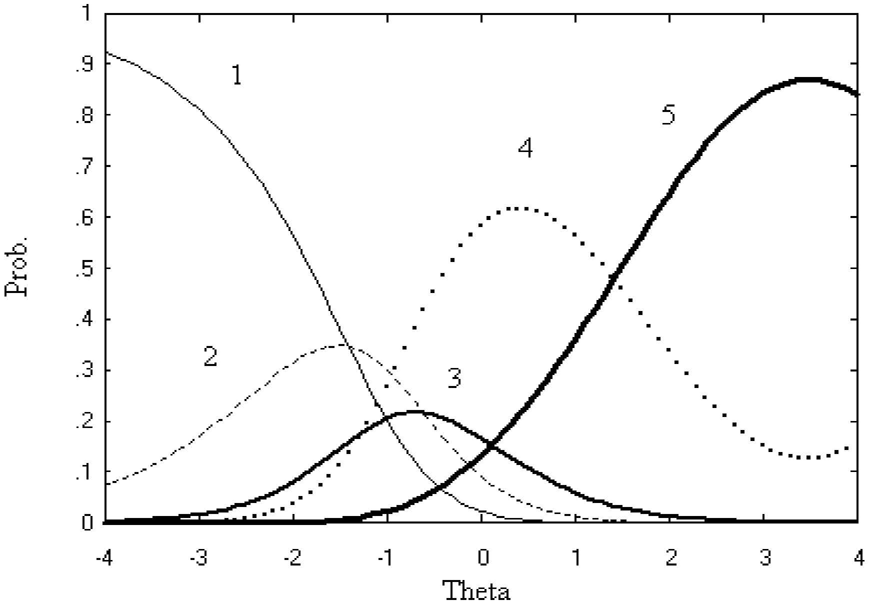
*Category Probability Plot for Item 21 with GGUM: Sample 5, Leadership Scale*



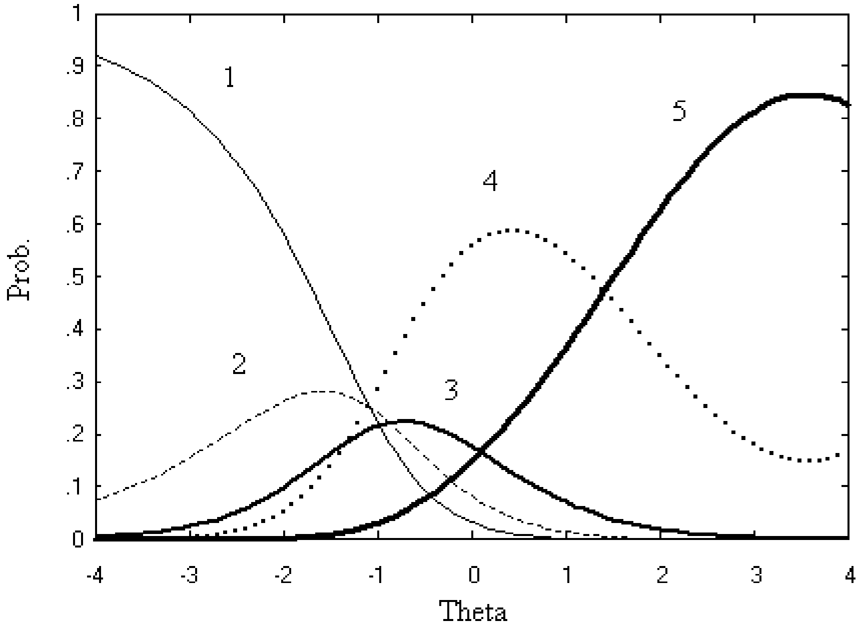
*Category Probability Plot for Item 21 with GGUM: Sample 6, Leadership Scale*



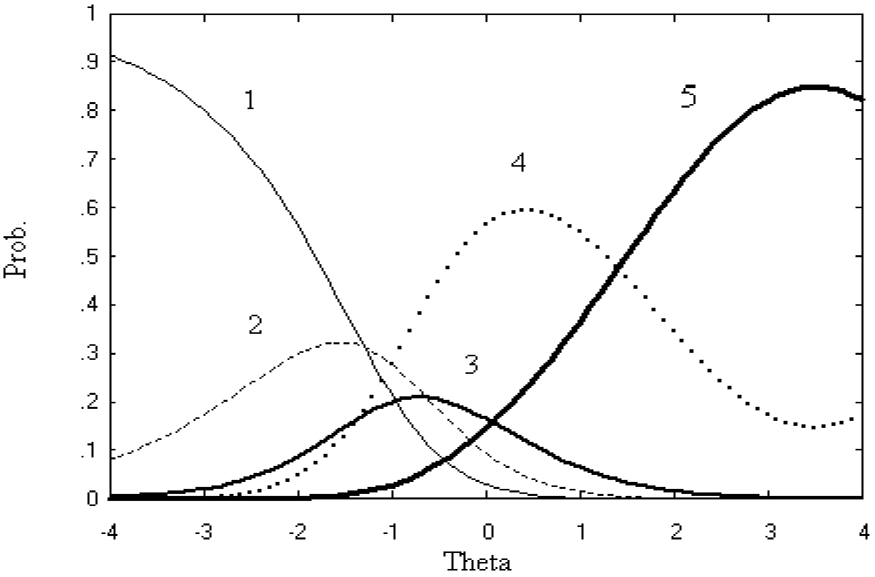
*Category Probability Plot for Item 21 with GGUM: Sample 7, Leadership Scale*



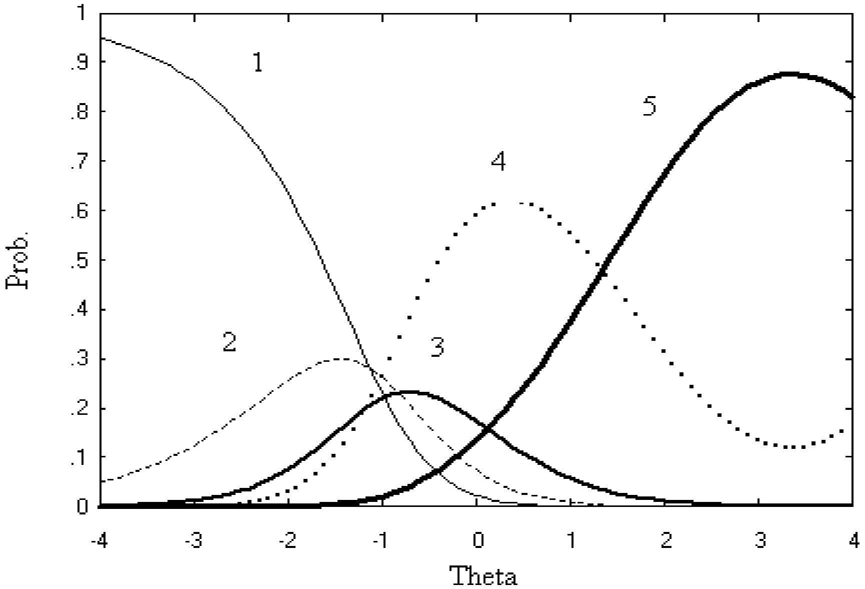
*Category Probability Plot for Item 21 with GGUM: Sample 8, Leadership Scale*



*Category Probability Plot for Item 21 with GGUM: Sample 9, Leadership Scale*



*Category Probability Plot for Item 21 with GGUM: Sample 10, Leadership Scale*



# REFERENCES

Ackerman, T. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*, 67-91.

Ackerman, T. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education, 7,* 255-278.

Ackerman, T. (1996). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement, 20*, 311-329.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716-723.

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

Andrich, D. (1988). The application of an unfolding model of the PIRT type to the measurement of attitude. *Applied Psychological Measurement, 12,* 33-51.

Andrich, D. (1995). Hyperbolic cosine latent trait models for unfolding direct responses and pairwise preferences. *Applied Psychological Measurement, 19*, 269-290.

Andrich, D. (1996). A hyperbolic cosine latent trait model for unfolding polytomous responses: Reconciling Thurstone and Likert methodologies. *British Journal of Mathematical and Statistical Psychology, 49*, 347-365.

Andrich, D., & Luo, G. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. *Applied Psychological Measurement, 17*, 253-276.

Andrich, D., & Styles, I. (1998). The structural relationship between attitude and behavior statements from the unfolding perspective. *Psychological Methods, 3*, 454-469.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

Bock, D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443-459.

Bockenholt, U., & Bockenholt, I. (1991). Constrained latent class analysis: Simultaneous classification and scaling of discrete choice data. *Psychometrika, 56*, 699-716.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*, 345-370.

Center for Teaching Quality. (n.d.). *Teacher Working Conditions Toolkit*. Retrieved September 2008, from http://www.teacherworkingconditions.org/dataanalysis/index.html.

Chernyshenko, O., Stark, S., Chan, K., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research, 36,* 523-562.

Chernyshenko, O., Stark, S., Drasgow, F., & Roberts, B. (2007). Constructing personality scales under the assumptions of an ideal point response process: Toward increasing the flexibility of personality measures. *Psychological Assessment, 19,* 88-106.

Cliff, N., Collins, L., Zatkin, J., Gallipeau, D., & McCormick, D. (1988). An ordinal I scaling method for questionnaire and other ordinal I data. *Applied Psychological Measurement, 12*, 83-97.

Coombs, C. (1950). Psychological scaling without a unit of measurement. *Psychological Review, 57*, 145-158.

Coombs, C. (1964). *A theory of data*. New York: Wiley.

Coombs, H., & Avrunin, G. (1977). Single-peaked functions and the theory of preference. *Psychological Review, 84*, 216-230.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth.

Davison, M. (1977). On a metric, unidimensional unfolding model for attitudinal and developmental data. *Psychometrika, 42*, 523-548.

De la Torre, J., Stark, S., & Chernyshenko, O. (2006). Markov chain monte carlo estimation of item parameters for the generalized graded unfolding model. *Applied Psychological Measurement, 30*, 216-232.

DeMars, C. (2004). Type I error rates for generalized graded unfolding model fit indices. *Applied Psychological Measurement, 28,* 48-71.

DeMars, C., & Erwin, T. (2003). Revising the scale of intellectual development: Application of an unfolding model. *Journal of College Student Development, 44,* 168-184.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B, 39*, 1-38.

DeSarbo, W., & Hoffman, D. (1986). Simple and weighted unfolding models for the spatial representation of binary choice data. *Applied Psychological Measurement, 10*, 247-264.

DeSarbo, W., Park, J., & Scott, C. (2008). A model-based approach for visualizing the dimensional structure of ordered successive categories preference data. *Psychometrika, 73,* 1-20.

du Toit, M. (Ed.). (2003). IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT. Lincolnwood, IL: Scientific Software International, Inc.

Ferguson, L. (1941). A study of the Likert technique of attitude scale construction. *The Journal of Social Psychology, 13,* 51-57.

Geisser, S., & Eddy, W. (1979). A predictive approach to model selection. *Journal of the American Statistical Association, 74*, 153-160.

Gelfand, A., & Dey, D. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society, B*, *56*, 501-514.

Haberman, S. (2006). Bias in estimation of misclassification rates. *Psychometrika, 71*, 387-394.

Habing, B., Finch, H., & Roberts, J. (2005). A Q3 statistic for unfolding item response theory models: Assessment of unidimensionality with two factors and simple structure. *Applied Psychological Measurement*, *29,* 457-471.

Hambleton, R., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston, MA: Kluwer-Nijhoff.

Hambleton, R., Swaminathan, H., & Rogers, J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hardouin, J., & Mesbah, M. (2004). Clustering binary variables in subscales using an extended Rasch model and Akaike information criterion. *Communications in Statistics: Theory and Methods, 33*, 1277-1294.

Hirsch, E., Emerick, S., Church, K., & Fuller, E. (2006). *Teacher working conditions are student learning conditions: A report on the 2006 North Carolina Teacher Working Conditions Survey*. Retrieved July 2007, from the Center for Teaching Quality’s Web site at http://www.teachingquality.org/pdfs/twcnc2006.pdf.

Hoijtink, H. (1990). A latent trait model for dichotomous choice data. *Psychometrika, 55,* 641-656.

Hoijtink, H. (1991). The measurement of latent traits by proximity items. *Applied Psychological Measurement, 15,* 153-169.

Houseman, E., Coull, B., & Betensky, R. (2006). Feature-specific penalized latent class analysis for genomic data. *Biometrics, 62,* 1062-1070.

Houseman, E., Marsit, C., Karagas, M., & Ryan, L. (2007). Penalized item response theory models: Application to epigenetic alterations in bladder cancer. *Biometrics, 63,* 1269-1277.

Jannarone, R. (1997). Models for locally dependent responses: Conjunctive item response theory. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 465-479). New York: Springer.

Johnson, M., & Junker, B. (2003). Using data augmentation and Markov chain monte carlo for the estimation of unfolding response models. *Journal of Educational and Behavioral Statistics, 28,* 195-230.

Joreskog, K. G., & Sorbom, D. (2006). *LISREL 8.8*. Lincolnwood, IL: Scientific Software International, Inc.

Kang, T & Chen, T. (2008). Performance of the generalized S-$X^2$ item fit index for polytomous IRT models. *Journal of Educational Measurement, 45,* 391-406.

Kang, T, & Cohen, A. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement, 31,* 331-358.

Kang, T., Cohen, A., & Sung, H. (2005, March). *IRT model selection methods for polytomous items*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Kline, R. (2005). *Principals and practice of structural equation modeling* (2nd ed.). New York, NY: The Guilford Press.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology,140,* 5-53.

Lin, T., & Dayton, C. (1997). Model selection information criteria for non-nested latent class models. *Journal of Educational and Behavioral Statistics, 22,* 249-264.

Luo, G. (1998). A general formulation for unidimensional unfolding and pairwise preference models: Making explicit the latitude of acceptance. Journal *of Mathematical Psychology, 42,* 400-417.

Luo, G. (2000). A joint maximum likelihood estimation procedure for the hyperbolic cosine model for single-stimulus responses. *Applied Psychological Measurement, 24*, 33–49.

Lou, G., Andrich, D., & Styles, I. (1998). The JML estimation of the generalized unfolding model incorporating the latitude of acceptance parameter. *Australian Journal of Psychology, 50*(3), 187-198.

Luecht, R., & Miller, T. (1992). Unidimensional calibrations and interpretations of composite traits for multidimensional tests. *Applied Psychological Measurement, 16,* 279-293.

Maraun, M., & Rossi, N. (2001). The extra-factor phenomenon revisited: Unidimensional unfolding as quadratic factor analysis. *Applied Psychological Measurement, 25,* 77-87.

McDonald, R., & Mok, M. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research, 30,* 23-40.

Meijer, R., & Baneke, J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Methods, 9,* 354-368.

Mislevy, R., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55,* 195-215.

Mokken, R. (1997). Nonparametric models for dichotomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 351-367). New York: Springer.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159-176.

Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement, 17,* 351-363.

Muraki E., & Bock, R. D. (1997). *PARSCALE 3: IRT based test scoring and item analysis for graded items and rating scales*. Chicago: Scientific Software International, Inc.

Nandakumar, R., Hotchkiss, L., & Roberts, J. (2002, April). *Attitudinal data: Dimensionality and start values for estimating item parameters*. Paper presented at the 2002 annual meeting of the American Educational Research Association, New Orleans, LA.

North Carolina Professional Teaching Standards Commission. (2006, December). Annual report from the North Carolina professional teaching standards commission. Retrieved April 13, 2008, from http://www.ncptsc.org/Annual%20Reports/NCPTSC%20Annual%20Report%20-%20submitted%20to%20state%20board.doc

Noel, Y. (1999). Recovering unimodal latent patterns of change by unfolding analysis: Application to smoking cessation. *Psychological Methods, 4*, 173-191.

Ostini, R., & Nering, M. (2006). Polytomous item response theory models. *In T.F. Liao (Series Ed.), Quantitative Applications in the Social Sciences: Series Number 07-144*. CA: Sage.

Reckase, M. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9*, 401-412.

Roberts, J. (2003, April). *An item fit statistic based on pseudocounts from the generalized graded unfolding model: A preliminary report*. Paper presented at the 2003 annual meeting of the American Educational Research Association, Chicago, IL.

Roberts, J., & Shim, H. (2008). *GGUM2004 Technical Reference Manual* (Georgia Institute of technology). Retrieved from http://www.psychology.gatech.edu.unfolding/FreeSoftware.html.

Roberts, J., Donoghue, J., & Laughlin, J. (1998). *The generalized graded unfolding model: A general parametric item response model for unfolding graded responses.* (Research Rep. RR-98-32). Princeton, NJ: Educational Testing Service.

Roberts, J., Donoghue, J., & Laughlin, J. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24*, 3-32.

Roberts, J., Donoghue, J., & Laughlin, J. (2002). Characteristics of MML/EAP parameter estimates in the generalized graded unfolding model. *Applied Psychological Measurement, 26*, 192-207.

Roberts, J., Fang, H., Cui, W., & Wang, Y. (2006). GGUM2004: A Windows-based program to estimate parameters of the generalized graded unfolding model. *Applied Psychological Measurement, 30*, 64-65.

Roberts, J., & Laughlin, J. (1996). A unidimensional item response model for unfolding responses from a graded disagree-agree response scale. *Applied Psychological Measurement, 20*, 231-255.

Roberts, J., Laughlin, J., & Wedell, D. (1999). Validity issues in the Likert and Thurstone approaches to attitude measurement. *Educational and Psychological Measurement, 59*, 211-233.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometric Monograph, No. 17.

Statistical Analysis Software (SAS). Cary, North Carolina: SAS is a registered trademark of SAS Institute Inc. in the USA and in other countries, 2002.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461-464.

Stark, S., Chernyshenko, O., Drasgow, F., & Williams, B. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology, 91*, 25-39.

Takane, Y. (1996). An item response model for multidimensional analysis of multiple-choice data. *Behaviormetrika, 23*, 153-167.

Takane, Y. (1998). Choice model analysis of the "pick any/n" type of binary data. *Japanese Psychological Research, 40,* 31-39.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*(4), 567-577.

Thurstone, L. (1927). A law of comparative judgment. *Psychological Review, 34*, 278-286.

Thurstone, L. (1928). Attitudes can be measured. *The American Journal of Sociology, 33*, 529-554.

Touloumtzoglou, J. (1999). Resolving binary responses to the visual arts attitude scale with the hyperbolic cosine model. *International Education Journal, 1*(2), 94-116.

van der Linden, W., & Hambleton, R. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer.

van Schuur, W., & Kiers, H. (1994). Why factor analysis often is the incorrect model for analyzing bipolar concepts, and what model to use instead. *Applied Psychological Measurement, 18,* 97-110.

von Davier, M., & Yamamato, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial-credit model. *Applied Psychological Measurement, 28,* 389-406.

Weekers, A., & Meijer, R. (2008). Scaling response processes on personality items using unfolding and dominance models. *European Journal of Psychological Assessment, 24*(1), 65-77.

Yen, W. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*, 245-262.

Yen, W. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-145.

Yen, W. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-213.

Yen, W., & Fitzpatrick., A. (2006). Item response theory. In R.L. Brennan (Ed.). *Educational Measurement* (4th ed., pp. 187-220). Westport, CT: Praeger.