To all of those people who believed in me

when most people thought that I wouldn't even finish college…

in particular,

Dr. Trudy B. Cunningham,

Dr. F. William Koko,

Dr. Robert M. Midkiff, and

Dr. Debra A. Mathinos

ACKNOWLEDGEMENTS

I'd like to express my sincere appreciation to Dr. William E. Loadman and Dr. Donald E. Haefele for their assistance in this piece as well throughout my experience at Ohio State. I'd also like to thank Brett Barnard for his data collection skills. Furthermore, I'd like for Dr. John Kennedy to know how much I treasure his vast knowledge and his charismatic personality. Everyone at The National Council of State Boards of Nursing deserves acknowledgment for their support. In particular, my gratitude goes out to Susan Woodward for her assistance in editing this document. A quick thanks goes out to Rebecca Goldstein for continually inspiring me to keep a 'qualitative' perspective on this mostly quantitative work. Lastly, this work would not have been possible if not for the unending patience, understanding, knowledge, flexibility, and complete faith that three people had in me. Thank you Ellen, mom, and dad.

VITA

December 29, 1971……..	Born, Somerville, New Jersey
1994…….	B. A., Bucknell University, Lewisburg, Pennsylvania
1994-1995……	Graduate Research Assistant, Teen Sexuality and Pregnancy Prevention Evaluation Team, Columbus, Ohio
1995-Present……	Psychometric Research Assistant, The National Council of State Boards of Nursing, Chicago, Illinois

PRESENTATIONS

Gorham, J. L., & Bontempo, B. D. (1996). Repeater Patterns on NCLEX™ using CAT versus NCLEX™ using Paper-and-Pencil Testing. Paper presented at the Annual Meeting of the American Educational Research Association, New York.

Julian, E. J., & Bontempo, B. D. (1996). Investigation into Decision Rules for NCLEX™ Candidates Who Run Out of Time. Paper presented at the Annual Meeting of the American Educational Research Association, New York.

FIELDS OF STUDY

Major Field: Education

TABLE OF CONTENTS

PAGE

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

FIGURE PAGE

INTRODUCTION

Introduction

In order for a construct to be scientifically useful, it must possess a constitutive definition, or in today's terms, a qualitative definition (Torgerson, 1958; and Margenau, 1950). This definition links an abstract construct to well known words and concepts. By having a clear constitutive definition, research questions can be formulated that can relate different constructs together. However, qualitative definitions of constructs exist only in the mind and language of the researcher, making anything more than metaphysical research unfathomable.

In order for scientific research to take place, the constructs must be operationalized (Kerlinger, 1986). This process translates the constructs' constitutive definitions into accessible, observable, and usable definition. By having clear operational definitions, hypotheses can be created and tested to see how relationships between operationally defined constructs exist in the real world. Once these relationships are established, inferences may be made that allow a researcher to pose possible answers to basic research questions. In essence, these inferences are made from the operational level back to the qualitative level.

Most operational definitions are an imperfect representation of the constitutive definition of the construct. However, there are times when the qualitative and

operational definition are not related at all. In these circumstances, even if the conclusions drawn from the hypotheses are accurate, the inferences made to the basic research questions may misrepresent the nature of a construct's relationship to the phenomena under investigation. Therefore, it is of utmost importance to ensure that the operational definition of a construct is in line with the underlying constitutive definition.

One important characteristic of a construct's operational definition is its level of measurement. The level of measurement of a given construct should follow from the qualitative definition of that construct. However, in today's world of quantification, constructs with nominal levels of measurement are often operationalized into ordinal, interval, or ratio levels of measurement.

In operationally defining a construct, a researcher will often use the results derived from a "tool." Tools take on different shapes, forms, sizes, and have different intents. The tool is often a written survey with various questions that are designed to be related to some given construct. When such a survey is assembled properly, subjects who possess a construct differently, respond differently to questions on the survey.

Item Response Theory

There are many ways in which a researcher can derive useful results from a properly constructed survey. The quantitative method that holds hegemony in psychometrics today is Item Response Theory (IRT). In IRT, researchers in the fields of statistics and social science methodology have developed many probabilistic models which allow a researcher to predict the way in which subjects will respond to a given

tool based on how they possess a given latent construct. In using IRT to derive results from a survey, a researcher employs a probabilistic model which best 'fits' the situation. Once the proper model has been chosen, one can determine how different subjects possess the given latent construct and how the survey questions relate to each other.

Traditionally, IRT has only acknowledged models that measure continuous quantities. These models are called latent trait models. The results derived from this type of model can be placed along a unidimensional continuum sometimes called a scale. One accepted latent trait model, is the Rasch model (Rasch, 1980). This model is a logistic model which allows the researcher to create a scale that is independent of the items and subjects used in the creation and application of that scale (Wright & Stone, 1977).

"In itself, however, the descriptive adequacy of these (latent trait) models need not imply that the continuous latent skills they posit accurately represent psychological reality" (Haertel, 1990). That is, there are times when the best operationalization of a construct may not be as an ordinal or interval variable because the construct may not be possessed to varying degrees within subjects. Instead, a subject might be classified as being one of several different latent types, those classifications being exhaustive, mutually exclusive, and having the ability to accurately describe a construct's existence within individuals of that classification. The type of models that are used to fit such constructs are called latent class models (LCM) (Lazarsfeld & Henry, 1968). Although latent class models are not typically thought of as IRT models, they do fit within the earlier definition of IRT.

Purpose

When one does not know the "psychological reality" of the construct under study, the choice of which type of model to use becomes a relevant issue. This is also true if the results of a survey are to be derived without knowledge of the level of measurement of the latent construct. This study will examine this issue by addressing the same question four times: which model, a Latent Class model or the Rasch model, is the most appropriate measurement model for each of four different sections of a survey in which the nature of the latent constructs is unknown. The four sections of the survey under investigation (three separate sections of a survey as well as the combination of those three sections) are parts of a written questionnaire that was designed to probe teacher's and principal's beliefs about teacher evaluation practices (Barnard, and Haefele, 1993).

LITERATURE REVIEW

Evolution of Probabilistic Models

Research in probabilistic models has evolved via two separate paths. Latent trait models were developed for educational and psychological studies and have become the mainstay for the field of psychometrics. On the other hand, latent class models came about through an outgrowth in sociological cross-table analysis and log-linear modeling (Langeheine & Rost, 1988). Although the two types of research practices are very similar in mathematical nature and use, they tend to be separated by the natural and artificial boundaries between the sociological and psychological fields. It is because of this division that this review will take two courses.

Latent Trait Models

There have been two separate tracks of development within the latent trait model paradigm. The first, from the educational field, was started as an attempt to better measure what Cronbach (1984) called measures of maximum performance (i.e., ability, aptitude, achievement, knowledge, or skills). The second, and more pertinent to this study, was primarily driven by the various fields within psychology that were aimed at measuring attitudes and interests. Unlike the developmental courses of latent trait and latent class models, these two tracks developed side by side and significant developments in one generally facilitated progress in the other as well.

Educational Measurement

There have been many developments in the history of measuring aptitudes and intelligence since mental measurement was first used in ancient China. Bypassing many early accomplishments, including Classical Test Theory, brings one to the evolution of Item Response Theory. Although advancements in IRT began in the 1940's with the accomplishments of individuals like Lawley (1943) and Tucker (1946), who first developed the Item Characteristic Curve (ICC), IRT lay dormant while classical test theory carried hegemony until Fred Lord pioneered IRT throughout the 1950's (Lord 1952, 1953a, 1953b). Developments following Lord's work can be channeled into two broad categories, the development of new measurement models and new procedures for estimating model parameters.

Many different IRT models have been developed over the years. All of these are transformations of Spearman's general factor model used in common factor analysis (McDonald, 1982). The oldest measurement models were only applicable to dichotomous items and took on two forms, those based on the normal ogive and those on the logistic ogive. Lord (1952) proposed the normal models which were comprised of one, two, and three parameters. The logistic models fall into these same three basic categories: the one parameter logistic (1PL) model or Rasch model (Rasch, 1980), the two parameter logistic model (2PL) (Birnbaum, 1968), and the three parameter logistic model (3PL) (Birnbaum, 1968). Three parameter models specify the probability of a correct (or incorrect) response to an item as a function of three parameters: item

difficulty, item discrimination, and a guessing parameter. Starting with the most inclusive, the 3PL includes all three parameters, while the 2PL excludes the guessing parameter. The Rasch model is the most exclusive, eliminating the guessing parameter and holding the item discrimination parameter to be a constant (De Ayala, 1993). There has been much debate over which model is superior, mostly between the 3PL and Rasch model. These debates have fueled many philosophical discussions over the nature of guessing on tests, as well as the ability to achieve "item free, sample free statistics" (Wright, 1991, 1992).

Later, these measurement models were expanded to allow for polytomous items. A few of the notable polytomous models include the rating scale model (Andrich, 1978), the partial credit model (Masters, 1982), the nominal response model (Bock, 1972), and the graded response model (Samejima, 1969).

In using probabilistic models, after a specific measurement model has been chosen, the next step is to estimate the model parameters. Simply stated, the goal is to analyze examinee responses to a tool, using a statistical function to estimate what is not known, i.e., the examinees' ability and items' characteristics. Birnbaum (1958) was the first to use the maximum likelihood (ML) function for estimation. Many slight and specific alterations to the ML function have been developed and put to use (Goodman, 1974; Bock & Aitken, 1981; Thissen, 1982a). Bayesian estimation was discovered to be useful (Samejima, 1969; Owen, 1975) for situations where maximum likelihood functions were unable to successfully estimate ability such as those for perfect scores, null scores, or tests with very few questions. Although extremely useful in these

situations, these estimation procedures are dependent upon an a priori belief about an examinee's ability (Hambleton & Swaminathan, 1985).

All of the above mentioned functions are non-linear and require the use of numerical iterative procedures to derive a solution. The Demming-Stephan (1940) and Newton-Raphson procedures are the two most common. With the developments in computer technology many algorithms have been developed to run these iterative procedures (Mislevy & Bock, 1982; Thissen, 1982b; Wingersky, Barton, & Lord, 1982; Wright & Linacre, 1991).

Attitude Measurement

Early developments in attitude measurement can be broken down and classified according to the type of variation produced by the tool. The three types of variation are stimulus centered, subject centered, and response (Torgerson, 1958; Dawis, 1987).

Stimulus centered latent scales, also called judgment scales (Dawis, 1987), ask respondents to rank a given set of stimulus or items. Here, the dimension of variation is in the items. Thurstone's method of paired comparisons (1929) and Stephenson's Q-sort technique (1953) are examples of methods devised to create such attitude scales.

Subject centered scales, also called individual differences scales (Dawis, 1987), require subjects to choose from a series of predetermined ordered response options, or rating scale, for each stimulus or item. In this situation, the variation in responses occurs along the rating scale's dimension. Likert's method (1932) and the semantic differential

(Osgood, Suci, & Tannenbaum, 1957) are common methods used to create subject centered scales.

In response scales, variation in both item and rating scale dimensions is accounted for. It can be said that Guttman (1944) made the first advances in methods for creating this type of deterministic scale. Later developments to the Rasch model further expanded response scales' use for the measurement of attitudes (Andrich, 1978). In these models, the dichotomous response Rasch model is expanded to allow for polytomous responses, which are presumed to be ordered responses. Variation in both items and rating scales is synthesized into a latent dimension on which the items, with their rating scale, can be placed. Thurstone put attitude items on a scale, creating within subject variation. Likert put subjects' attitude on a scale, creating between subjects variation, and the Rasch model put them both on the same continuum, creating the first truly latent scale.

Latent Class Models

Green (1951) first used the term "latent class model," but the first systematic development of latent class models was done by Lazarsfeld (1950). In this model, the researcher need no more than nominal manifest categories (independent variables) to derive information about some latent variable (Lazarsfeld & Henry, 1968). "(A latent class model) identifies homogeneous subgroups that are characterized by their choice probabilities for a set of alternatives" (Bockenholt & Bockenholt, 1990). In many

respects, latent class analysis is the categorical counterpart to factor analysis in the continuous world (Everitt and Dunn, 1988).

Originally researchers were hesitant to use these models for several reasons. The early methods for parameter estimation, such as those developed by Anderson (1954) or Lazarsfeld and Dudman (1951), sometimes produced non-admissible probability estimates (greater than 0 or less than 1). McHugh (1956, 1958) first used the maximum likelihood function for estimation of LCMs which solved much of this problem. Still, like the early latent trait models, none of these methods could handle polytomous manifest variables or more than one latent variable. That is, until Goodman (1974), using a specific estimation maximization (EM) algorithm and ML estimation, showed how these models could be made applicable to polytomous manifest variables as well as multiple latent variables. Goodman also resolved one of the other early problems with latent class models. That is, he was able to constrain parameters of the model to fixed values, thereby enabling him to also set two or more parameters to be equal.

Many models were developed which were all shown to have mathematical similarities. For example, Formann (1984) illustrated how the constrained model proposed by Goodman was really only a specific case of his linear logistic LCA. Formann also demonstrated that his model was merely a specific case of Haberman's (1979) general LCM.

Mixture Distribution Models

A mixture distribution model (MDM) assumes that observed data come from a mixture of two or more latent populations as opposed to one homogeneous population (Everitt & Hand, 1981). "These subpopulations are only defined by their property of being homogeneous in the sense that a particular model with a specified set of parameters holds for these individuals. In particular, they are not defined by manifest variables like sex or age…" (von Davier, 1994). The only assumptions that are necessary are the number of subpopulations and how those subpopulations are to be modeled.

For example, it can be specified that for some set of data, two subpopulations (also called classes) exist. It can also be specified that one of those classes will be modeled using the Rasch model while the other class will be modeled via an LCM. If the data fit this hypothetical model, then we can say that variation in the responses from those individuals in Class 1 is due to some unidimensional measurable latent construct. A quantitative score is appropriate for describing these individuals. Variation in responses from those individuals in Class 2 is assumed to be random. When the classes are all modeled as latent classes, the type of MDM is called an LCA. When they are all modeled as Rasch models, the MDM is called Mixed Rasch Model (MIRA). And when there are some latent classes and some Rasch classes, the model is called a Hybrid Model (Yamamoto, 1989).

Mathematical Assumptions

Three main axioms are applied when using a latent trait model. Monotonicity is implied, meaning that the probability of a positive response increases as a function of the latent trait. The latent trait must be unidimensional, implying that the items and people must relate to each other on just one underlying dimension. And lastly, local stochastic independence must be met, requiring all items to be independent of one another. In addition, the statistics used by the Rasch model are sufficient statistics: the number of positive responses given by a subject is the only information necessary for describing the subject parameters and the number of positive responses to an item is the only information necessary for describing the item parameters (Rasch, 1980).

The basic assumptions of latent class models are less stringent to fulfill than those of latent trait models. The first assumption is that the latent classes in the population must be mutually exclusive and exhaustive. For each latent class, each indicator (item or manifest variable) must have a specific probability of occurrence. And lastly, for each latent class, the manifest variables much be locally independent (Langeheine, 1988).

The Mathematics Behind Latent Trait and Latent Class Models

For the dichotomous mixed-Rasch model, the probability of an item response x_{vi} can be expressed as follows:

$$P(\mathbf{x}_{vi}) = \sum_{g=1}^{G} \pi_g \left(\frac{e^{x_{vi}\left(\xi_{vg} - \sigma_{ig}\right)}}{1 + e^{\xi_{vg} + \sigma_{ig}}} \right). \quad (1)$$

For equation 1, ξ_{vg} is the ability of person v in group g, σ_{ig} is the item difficulty of item i in group g, and π_g is the relative class size (Rost, 1990). For the various polytomous models, the mixed-Rasch model is as follows:

$$P(\underline{x}) = \sum_{g=1}^{G} \pi_g \pi_{rg} \frac{\sum e^{\sum_{i=1}^{I} \sum_{s=1}^{x} \alpha_{ixg}}}{\gamma_{rg}}. \quad (2)$$

Here, $\underline{x}$ is the response pattern $\underline{x} = (x_1,\ldots,x_I)$ & $x_i \in \{0,\ldots,m\}$, π_{rg} is the probability of score r in class g, and γ_{rg} is the symmetric function of order r in class g (von Davier, 1994).

The general latent class model is assumed to be a linear function of some parameters which are dependent on variable i, the response category x, and the latent class g. This function can written as follows (von Davier, 1994):

$$f_{gix} = \log \frac{p_{vix}}{p_{vix-1}}. \quad (3)$$

In this case, the probability of person v choosing response category x on variable i can be defined as follows (von Davier, 1994):

$$P_{vix} = \sum_{g=1}^{G} \pi_g \frac{e^{\left(\sum_{s=0}^{x} f_{gis}\right)}}{\sum_{t=0}^{m} e^{\left(\sum_{s=0}^{t} f_{gis}\right)}}. \quad (4)$$

Model Comparison

In life, there are times when "the chicken or the egg" debate becomes a relevant analogy to a phenomena. Such is the case when latent trait models are used to model categorical latent variables and latent class models are used to model continuous latent variables.

A latent trait supporter might argue that latent class models do not have the ability to produce precise "measurements" since they do not have the ability to express a latent construct on a continuous level of representation. Latent class supporters would

argue that they do have this ability, for they can derive as many classes as there are subjects. If there were an infinite number of subjects, then there could be an infinite number of classes. In a sense, latent class models estimate continuousness by taking an infinite number of ordered classifications. However, the latent trait paradigm enables one to objectively judge the distance between points on a line. In the case of representing a continuous quantity, this is more defensible.

A latent class model supporter would argue that a latent trait model cannot "classify" people precisely since a unidimensional ruler has an infinite number of points. In classifying people using a latent trait model, the ruler is divided into discrete sections which are estimated from the places where the data cluster around certain points of the ruler. In a sense, the number of categories can never be exact since they are always estimated, as are the individual classifications. For this reason, the latent class paradigm seems more admissible in representing a categorical construct since it enables one to objectively classify subjects without having to somewhat arbitrarily estimate the categories and their quantitative boundaries.

Research has shown that the Rasch model can be formulated as a log-linear model (Clogg, 1988, & Kelderman, 1984). From Haberman's model (1979), it is also known that latent class models are also forms of the log-linear model (Langeheine & Rost, 1988). In essence, all possible latent trait and latent class models are derivations of the most general log-linear model.

In all, there has been extensive research done on the topics of latent trait and latent class models. Previous research demonstrated how to use latent trait models for

the measurement of continuous variables such as performance and attitude. Research has also shown how to use latent class models for the analysis of categorical constructs. However, the research is inconclusive regarding which model is most appropriate for situations where the level of measurement of the latent constructs is unknown. At present, no other criteria besides the level of measurement, is used to figure out which type of model, latent class or latent trait, is most appropriate. There is other available criteria, some of which is outlined in Chapter III. By employing these criteria in the decision process, a researcher can justify a type of models use without a priori knowledge of the level of measurement of the latent construct

METHODOLOGY

This chapter outlines four things. It provides a description of the sample as well as a description of the instrument. The specifications for each of models and the software WINMIRA are also included. Lastly, it describes the criteria that were used to evaluate the two different types of models.

Sample

A multiple stage sampling process was used to collect this data (Barnard & Heafele, 1993) A group of 61 school districts was systematically chosen from the 612 public school districts in the sate. Within these districts, one elementary and one high school were chosen at random. An additional high school and elementary school were also included for the two largest school districts. Within each school, the principal was asked to choose 4 teachers at random to administer the survey. Of the 115 principals and 460 teachers who agreed to participate, response from 433 (90.2%) teachers and 110 principals (96.6%) were submitted and used.

Instrument

Subjects were asked Likert-type questions in three different sections of a written questionnaire designed to probe opinions about teacher evaluation practices across the

state (Barnard & Haefele, 1993). A copy of these sections of the questionnaire is found in Appendix A. The first section of the questionnaire was called Content of the Evaluation. This section contained 15 examples of criteria for the evaluation. The available response options were never used (0), rarely used (1), moderately used (2), frequently used (3), and extensively used (4). The second section was titled Utilization of the Evaluation, which contained 11 questions and the same response options as the first. This section was aimed at measuring the extent to which the evaluative information was used. The final section, Attitudes Toward Evaluation, contained 25 statements about certain aspects of the evaluation process. Following each statement were five response options: strongly disagree (1), disagree (2), agree (3), strongly agree (4), and undecided (5).

Specification of Models

The data from each section of the questionnaire as well as the combination of all three sections were fit to 16 different models using the software WINMIRA (von Davier, 1994) (See Table 1). Since these models are not defined by independent variables such as position, the data from both teachers and principals were combined into one data set. Of

Table 1

The Number of Classes and Type of Distribution Specified by Each Model

	A Priori Number of Classes (or Subpopulations)									
Model #	1	2	3	4	5	6	7	8	9	10
1	LCA									
2	LCA	LCA								
3	LCA	LCA	LCA							
4	LCA	LCA	LCA	LCA						
5	LCA	LCA	LCA	LCA	LCA					
6	LCA	LCA	LCA	LCA	LCA	LCA				
7	LCA	LCA	LCA	LCA	LCA	LCA	LCA			
8	LCA	LCA	LCA	LCA	LCA	LCA	LCA	LCA		
9	LCA	LCA	LCA	LCA	LCA	LCA	LCA	LCA	LCA	
10	LCA	LCA	LCA	LCA	LCA	LCA	LCA	LCA	LCA	LCA
11	Rasch									
12	Rasch	Rasch								
13	Rasch	Rasch	Rasch							
14	Rasch	LCA								
15	Rasch	LCA	LCA							
16	Rasch	Rasch	LCA							

these sixteen models eleven are of direct interest. The first 10 models are all LCA only models, varying only in the number of classes they posit (1-10 classes). The ten class model was chosen arbitrarily as the model with the highest number of classes. Models with more than 10 classes are not parsimonious and would require strong a priori rationale for their use. The eleventh model is the ordinal Rasch model. The data were fit to the other five models as a demonstration of other possible MDMs. In all of these situations, the model parameters were assumed to be structural or fixed and were derived in WINMIRA using a CML (Conditional Maximum Likelihood) method. There were two conditions. The first was the specific measurement model to be used, in this case the ordinal model, also called the partial credit model (Masters, 1982), was used. This model did not require the item thresholds, to be the same for every item, and it did not require those thresholds to be equidistant from each other within any item. An item threshold is the minimum probability of choosing a particular response category for subjects of a given score. For the LCA only models, the thresholds were class specific meaning that the thresholds were different for every LCA latent class. The second CML condition was the latent distribution of scores which was fully parameterized in this situation meaning that there was one parameter for each score in each class. Subject scores were assumed to be incidental or random variables and were estimated using the UML (Unconditional Maximum Likelihood) procedure in WINMIRA after the model parameters were derived.

Criteria for Assessing Models

The primary criteria for evaluating the most appropriate model was goodness of fit statistics. Other criteria were parsimony and the effects of misfitting items. Also, two relatively unused criteria were employed. Item profiles analysis, as well as the distribution of latent classifications at various levels of the Rasch continuum were used. This process of choosing the most appropriate model is not exact. Information from all of these criteria was gathered and a best judgment on the part of the author was used to synthesize the information into conclusions.

Goodness of Fit

Model fit was assessed by reporting log-likelihood goodness of fit statistics, as well as Akakie and Bayesian information criteria (AIC and BIC respectively). These two criteria are calculated as follows: AIC= -2 (log likelihood) + 2 κ, and BIC= -2 (log likelihood) + (log N) (κ). In these two equations, N = number of subjects and κ = number of parameters in the model. κ is defined by each model in the following manner:

$$\text{LCA, } \kappa = [\, I\,(m-1)\,(G)\,] + (G-1); \tag{6}$$

$$\text{MIRA, } \kappa = [\, I\,(m-1)\,(2G)] + (2G+1); \text{ and} \tag{7}$$

$$\text{Hybrid, } \kappa = [\, I\,(m-1)\,(G_L)\,] + (G_L-1) + [\, I\,(m-1)\,(2G_R)] + (2G_R+1) + 1. \tag{8}$$

In equations 6, 7, and 8, I is the number of items, m is the number of response options, and G is the number of classes. In these equations, as the number of classes increases the number of model parameters increases and the log-likelihood decreases. As the data are fit to models with a higher number of classes, and thus a higher number of parameters, the better the data fit the model. As a result, the goodness of fit statistics decrease in magnitude. If all possible parameters are used the resulting log-likelihood for this saturated model is 0. It should be obviously that this saturated model which sorts N classes into N categories does provide a perfect fit but little useful information. The information criteria counteracts this by increasing the goodness of fit statistics using the number of parameters in the model.

Item-Q-Index

The Item Q-index was used to evaluate misfittting items on all Rasch models (Rost and von Davier, 1994). The index is applicable to dependent variables of any level of measurement and is therefore very useful for our purposes. Q is based on the likelihood of the observed response pattern for the item and is calculated as follows:

$$Q = \frac{\sum_{v}\left(x_{vi} - x_{v,opt}\right)\beta_{v}}{\sum_{v}\left(x_{v,pess} - x_{v,opt}\right)\beta_{v}} \qquad (9)$$

In equation 9, x_{vi} is the observed response pattern, $x_{v,\,opt}$ is the optimum pattern or the Guttman pattern, and $x_{v,\,pess}$ is the pessimum pattern or the anti-Guttman pattern. A perfect fitting item would have a Q of 0 while a misfitting item would have a Q of 1.

Item Profiles

An item profile was built for each section, based on the best fitting LCA-only model. An item profile is a plot of the conditional response probabilities of each item response for each class. If the response probabilities for two classes do not cross, then those classes are ordered and dimensionality is appropriate. If they do cross, then those classes are unordered, making nominality seem just.

Distribution of Latent Classifications

The best fitting LCA-only model was compared with the ordinary Rasch model (the one-class Rasch model). A distribution of subjects' score estimates based on the ordinary Rasch model was plotted and each subject's LCA class was noted. If subject's each class congregate around given sections of the distribution then some form of unidimensional latent continuum may be appropriate. If the subjects display random abilities, then dimensionality may be inappropriate.

RESULTS AND DISCUSSION

Section 1: Content of the Evaluation

Goodness of Fit

The goodness of fit statistics for each model are displayed in Table 2. The data fit the ordinary Rasch model containing 119 parameters better than the two-class LCA which had 121 parameters. To aid in choosing the best fitting LCA-only model, two plots were created. These plots were of the information criteria (See Figure 1) and component log-likelihood (See Figure 2). According to the Akakie information criteria (AIC), the three, four, or five class models all fit while the component log-likelihood adds support to this claim. The Bayesian information criteria (BIC) provides little additional information. The AIC of the four-class model was slightly better than the other two, thus the four-class LCA was chosen from the three best fitting models as the best fitting LCA. The frequency of responses to each possible option can be seen in Table 3, and each of the LCAs' relative class sizes can be seen in Table 4. A frequency distribution of estimated scores, or thetas, was created (See Figure 3). The mean

Table 2

Goodness of Fit Statistics for Content of the Evaluation

Model	*Classes*	*Parameters*	*Log-Likelihood*	*AIC*	*BIC*	*Iterations*	*Component Log-Likelihood*
LCA	1	60	-10504.23	21128.46	21385.17	33	0
LCA	2	121	-9423.02	19088.04	19605.74	36	1081.21
LCA	3	182	-8973.13	18310.25	19088.94	172	449.89
LCA	4	243	-8699.77	17885.54	18925.22	129	273.36
LCA	5	304	-8547.46	17702.92	19003.59	209	152.31
LCA	6	365	-8436.63	17603.26	19164.92	167	110.83
LCA	7	426	-8319.22	17490.44	19313.09	179	117.41
LCA	8	487	-8229.32	17432.63	19516.27	184	89.9
LCA	9	548	-8174.43	17444.86	19789.49	192	54.89
LCA	10	609	-8091.52	17401.04	20006.66	204	82.91
Rasch	1	119	-9103.12	18444.25	18953.39	193	
Rasch	2	237	-8578.43	17630.85	18644.86	250	
Rasch	3	355	-8370.66	17451.32	18970.19	250	
Mixed: Rasch/LCA	2	180	-8777.61	17915.22	18685.36	191	
Mixed: Rasch/LCA/LCA	3	241	-8636.26	17754.52	18785.64	250	
Mixed: Rasch/Rasch/LCA	3	298	-8479.53	17555.07	18830.07	250	

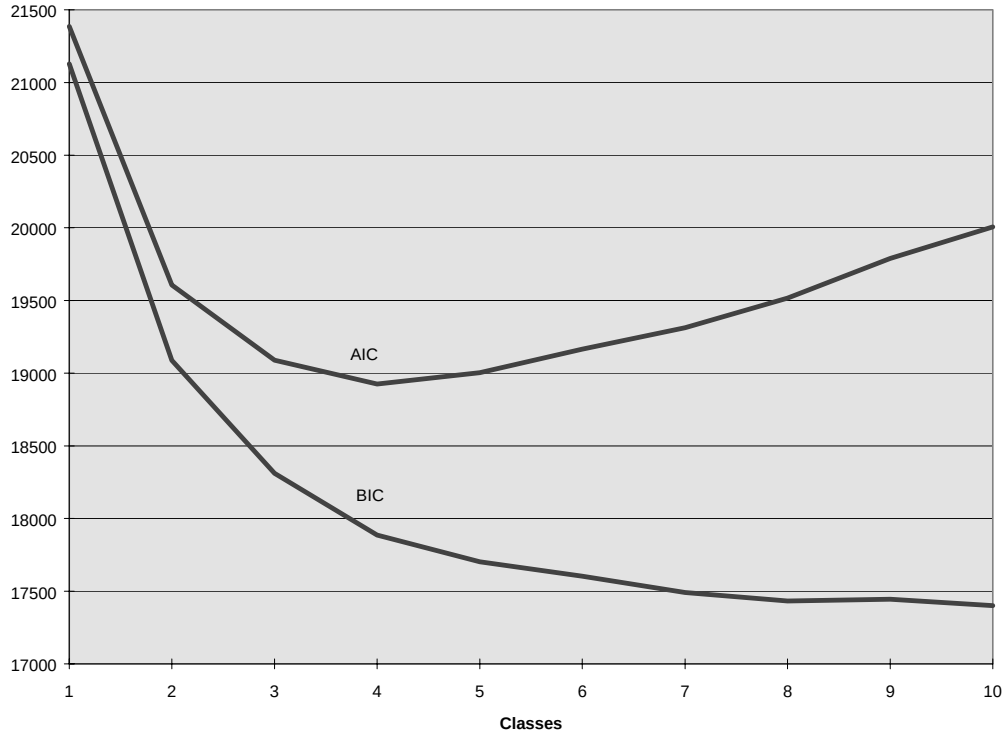

Figure 1. Information criteria for Content of the Evaluation.

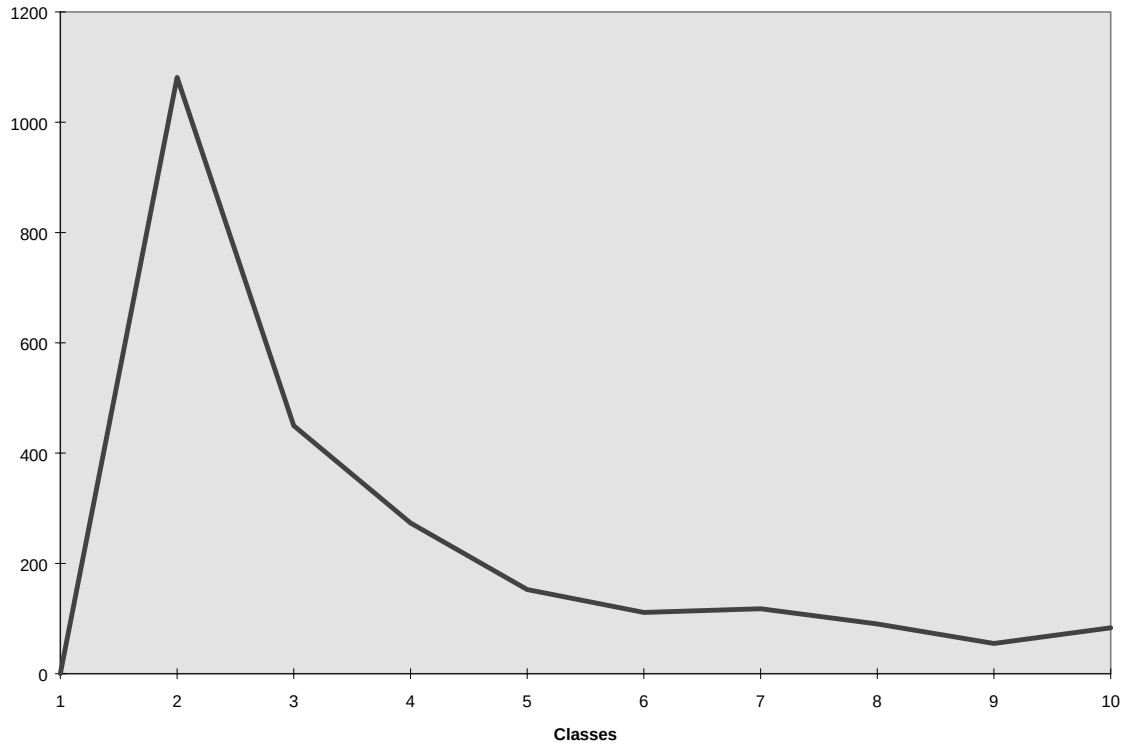

Figure 2. Component log-likelihood for Content of the Evaluation.

Table 3

Frequency of Subjects Using Each Response Option on Content of the Evaluation

Item	Response					
	0	1	2	3	4	Missing
1	0	38	86	196	191	22
2	14	15	72	211	219	2
3	47	110	173	138	63	2
4	11	14	51	211	245	1
5	15	22	90	210	193	3
6	15	15	55	199	246	3
7	12	32	107	204	174	4
8	10	5	68	196	251	3
9	10	19	122	214	164	4
10	60	93	147	149	81	3
11	27	50	138	197	116	5
12	10	6	38	197	281	1
13	28	38	101	207	152	7
14	49	80	154	153	92	5
15	13	38	156	205	117	4

Table 4

Relative Class Size for Content of the Evaluation

LCA 1	*LCA 2*	*LCA 3*	*LCA 4*	*LCA 5*	*LCA 6*	*LCA 7*	*LCA 8*	*LCA 9*	*LCA 10*
0.57641	0.42359								
0.50087	0.39852	0.10061							
0.42636	0.28355	0.22846	0.06163						
0.33493	0.26367	0.19820	0.16758	0.03561					
0.32522	0.23051	0.20696	0.15741	0.04427	0.03563				
0.26954	0.24239	0.16779	0.11912	0.09894	0.06505	0.03716			
0.26404	0.19803	0.16605	0.12731	0.09662	0.06469	0.04951	0.03374		
0.17550	0.16003	0.15189	0.14394	0.10238	0.10025	0.08924	0.04300	0.03376	
0.18349	0.16866	0.16706	0.14702	0.11617	0.05560	0.05102	0.04704	0.03580	0.02814

Rasch 1	*LCA 2*	*LCA 3*
0.80803	0.19197	
0.64808	0.24252	0.10940

Rasch 1	*Rasch 2*	*LCA 3*
0.57979	0.42021	
0.55291	0.31821	0.12888

Rasch 1	*Rasch 2*	*Rasch 3*
0.52234	0.29083	0.18684

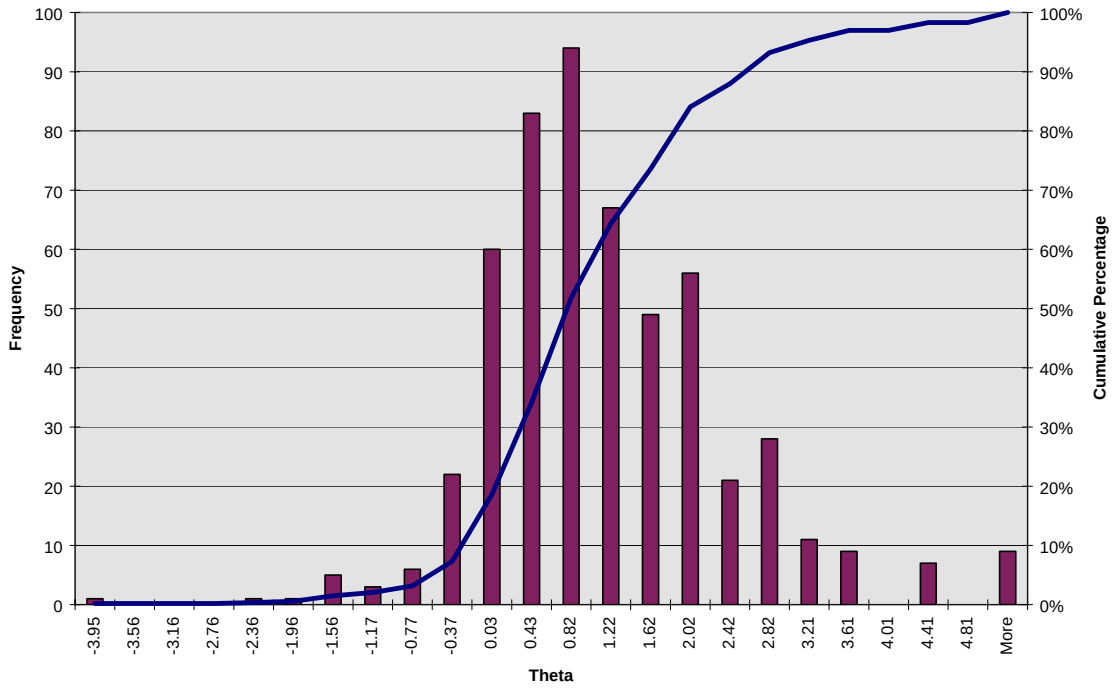

Figure 3. Frequency distribution of thetas for the ordinary Rasch model on Content of the Evaluation.

estimate was 1.0 with a standard deviation of 1.22. The raw score reliability calculated by the odd/even split-half method was $\alpha \geq 0.85$ (Cronbach, 1951). Based on the Item Q-index, none of the items showed severe misfit (See Table 5).

Item Profiles

The probability of response on each item for each class was plotted for the series of response options (Never Used... Extensively Used) (See Figures 4-8). At first glance, these item profiles may look like a scattered mess. However, what is discernible rather quickly is that Class 1 had a high proximity to response 3 (Frequently Used). Subjects in this class believe that their evaluations cover a wide range of areas. Class 2 had a high proximity to response 4 (Extensively Used) which indicates that the previous claim is even more true for subjects in this class. Class 3 leaned toward either response 1 or 2 (Rarely to Moderately Used), while Class 4 opted for response 0 (Never Used). Subjects in these two classes generally believe that their evaluations cover either none of the listed areas or possibly only one or two of the areas. Notice that within each response category and across items, the response probability of the class with the highest probability rarely drops below the response probabilities of the other classes.

Distribution of Latent Classifications

The distribution of classifications derived from the four-class LCA at various levels of the Rasch continuum is shown in Figure 9. This chart shows that the four

Table 5

Item-Q-Index for Content of the Evaluation

Item #	Q
1	0.1420
2	0.1108
3	0.2065
4	0.0903
5	0.0928
6	0.0797
7	0.0816
8	0.0809
9	0.1571
10	0.1484
11	0.1013
12	0.1098
13	0.1136
14	0.1637
15	0.1591

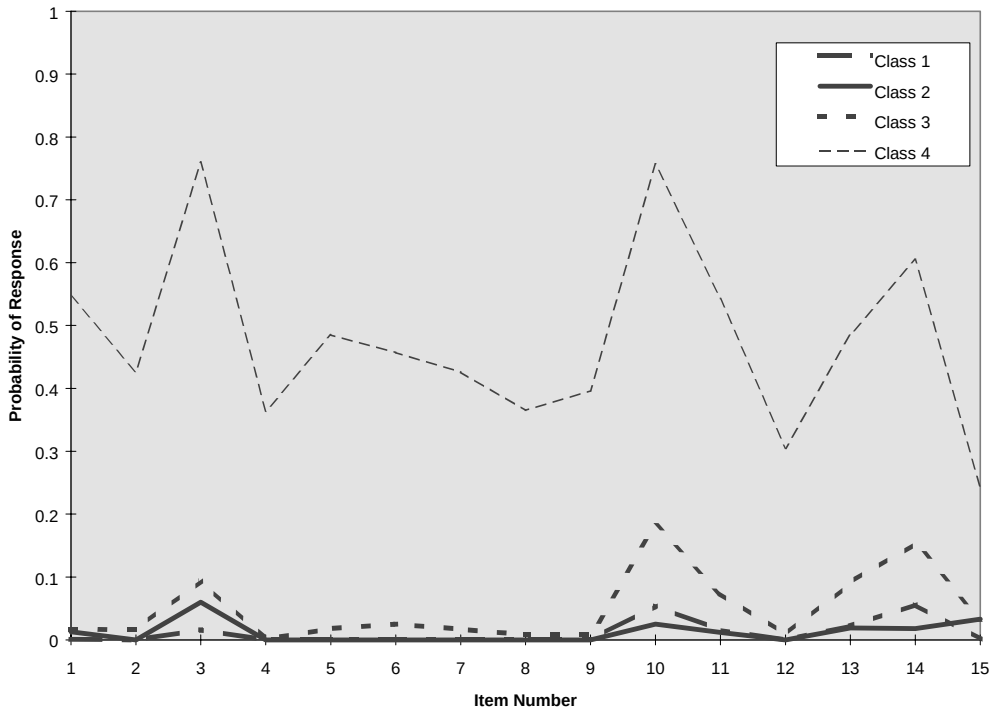

Figure 4. Each class's expected probability of responding with category 0 (Never Used) on Content of the Evaluation.

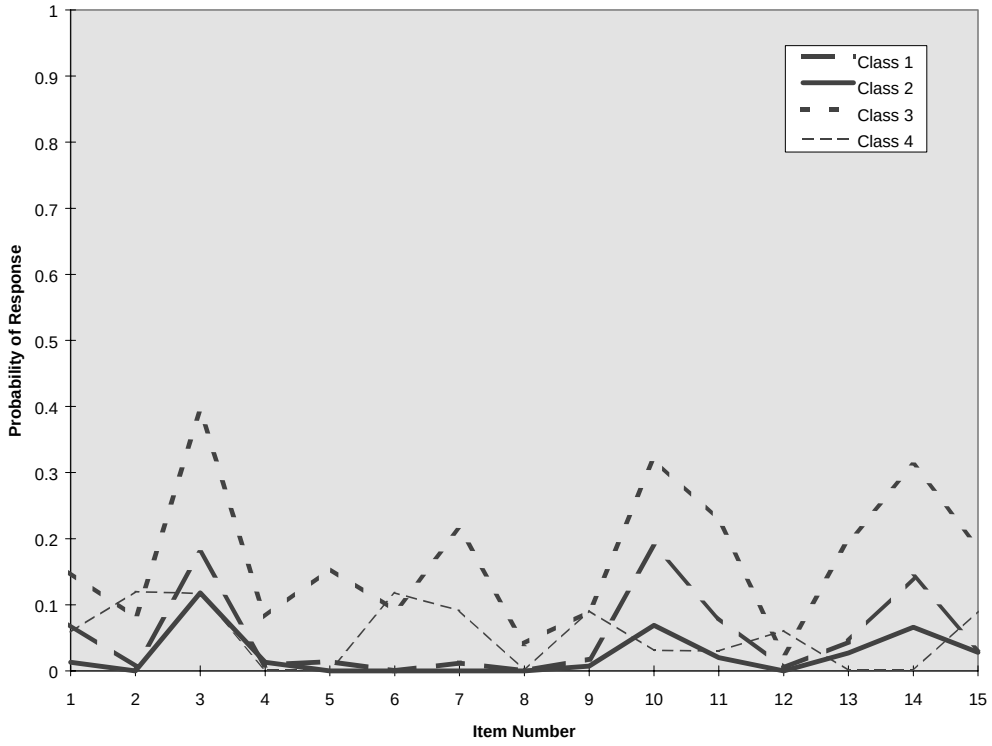

Figure 5. Each class's expected probability of responding with category 1 (Rarely Used) on Content of the Evaluation.

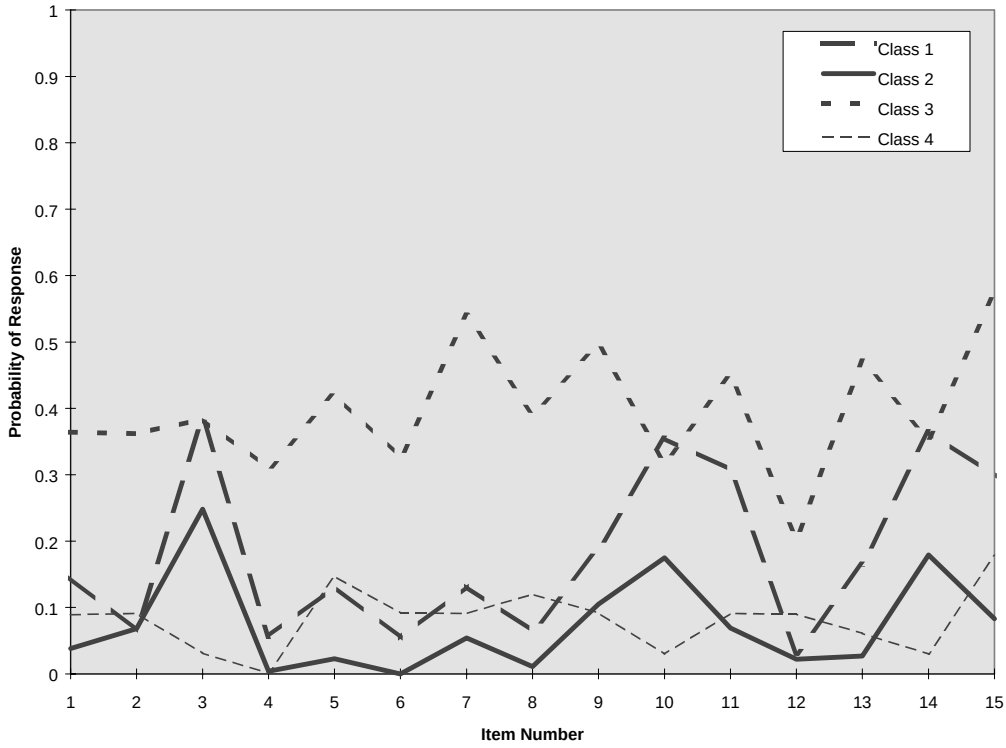

Figure 6. Each class's expected probability of responding with category 2 (Moderately Used) on Content of the Evaluation.

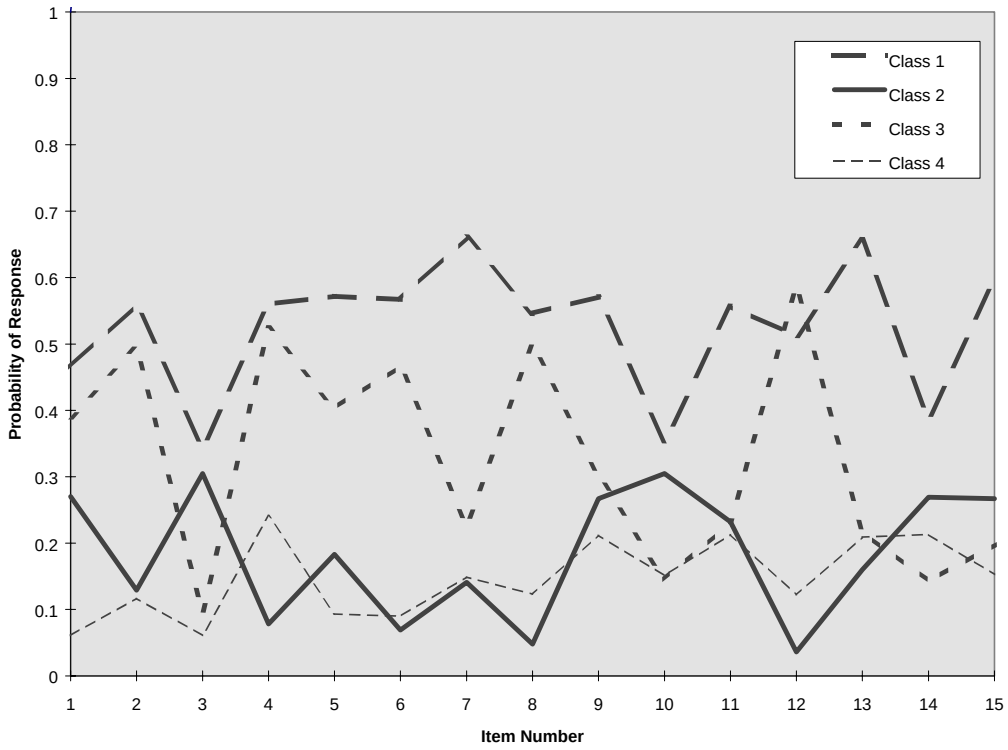

Figure 7. Each class's expected probability of responding with category 3 (Frequently Used) on Content of the Evaluation.

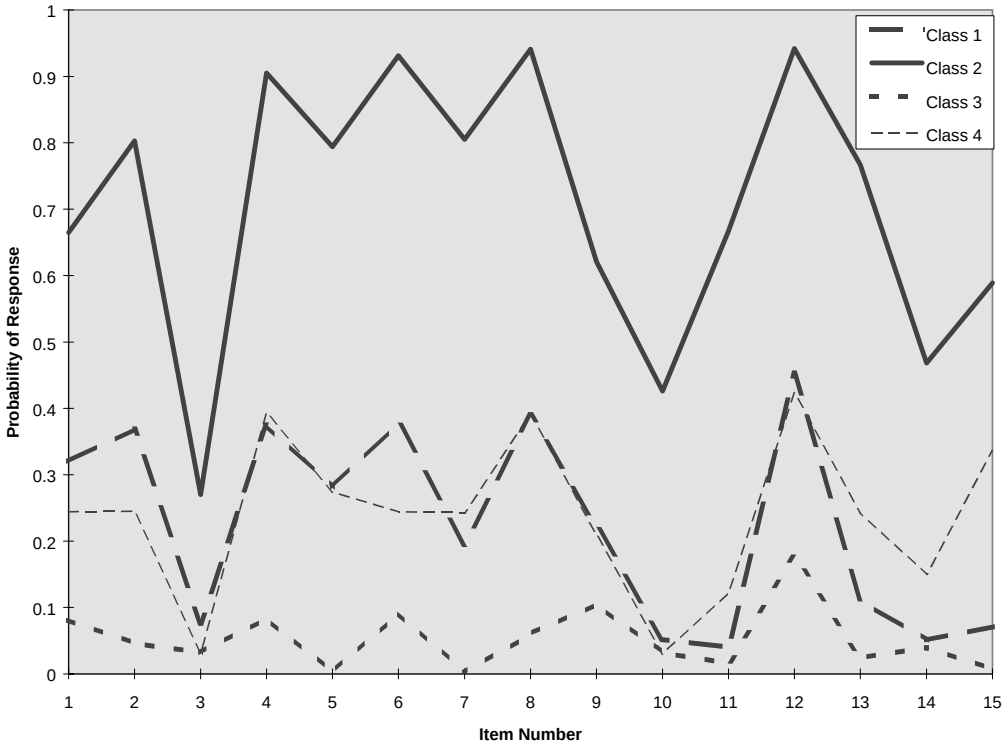

Figure 8. Each class's expected probability of responding with category 4 (Extensively Used) on Content of the Evaluation.

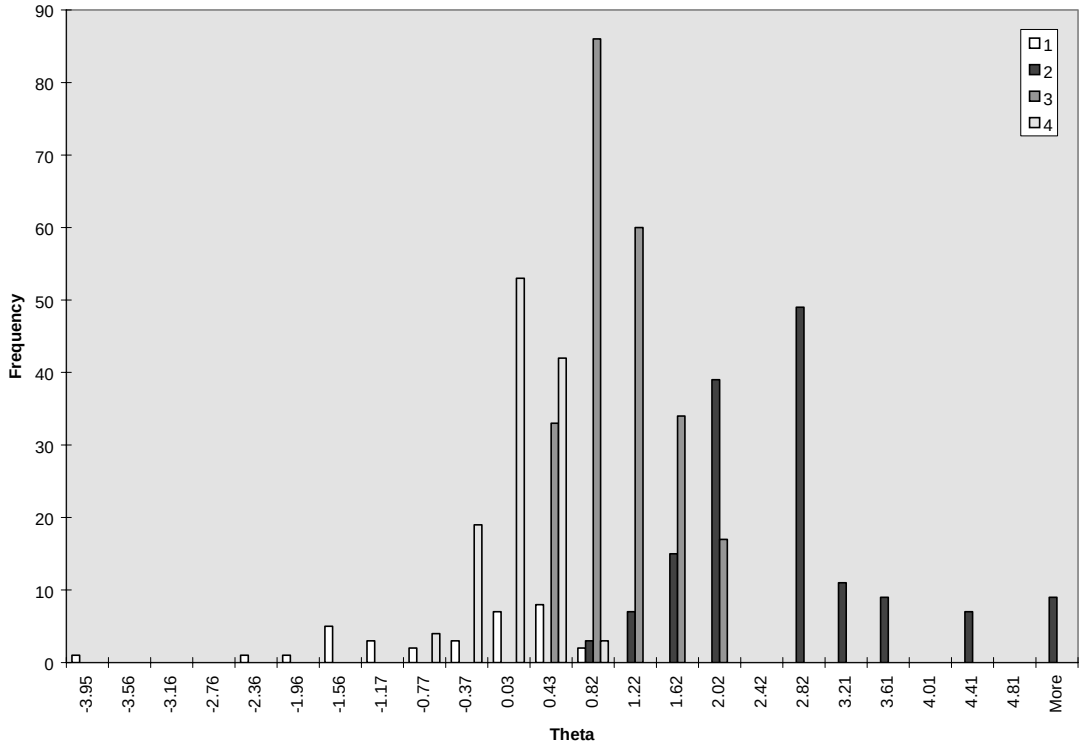

Figure 9. Distribution of subjects within each classification of the 4 Class LCA at various levels of estimated scores from the ordinary Rasch model on Content of the Evaluation.

classes had different means. Notice too, that these means were in close proximity to each other and contained plenty of overlap from one distribution to the next.

Discussion

In this section, a four-class ordinal LCA derives information about the latent variable that would be lost if only the Rasch score estimates were used in operationalizing the latent construct. The opposite is also true: The Rasch model derives information that would be lost if only a four-class LCA were employed. This later outlook is the more convincing of the two. Support for this comes from three sources. The fit of the Rasch model is better than that of two class LCA model which has a comparable number of parameters. The item profiles also revealed that there was little overlap in the probability of responding to a given category between the different classes. Lastly, the distribution of classifications revealed that the classes posses order. Overall, the Rasch model is most appropriate for this section. What's more, the latent construct underlying the content of subjects' evaluations has an interval level of measurement.

Section 2: Utility of the Evaluation

Goodness of Fit

The goodness of fit statistics for this section can be found in Table 6. The ordinary Rasch model had 87 parameters. The data fit this model better than the less

Table 6

Goodness of Fit Statistics for Utility of the Evaluation

Model	*Classes*	*Parameters*	*Log-Likelihood*	*AIC*	*BIC*	*Iterations*	*Component Log-Likelihood*
LCA	1	44	-8330.89	16749.78	16938.03	33	0
LCA	2	89	-7789.34	15756.69	16137.48	71	541.55
LCA	3	134	-7470.21	15208.43	15781.75	86	319.13
LCA	4	179	-7325.23	15008.47	15774.32	88	144.98
LCA	5	224	-7210.16	14868.32	15826.71	135	115.07
LCA	6	269	-7086.32	14710.64	15861.56	114	123.84
LCA	7	314	-7027.34	14682.69	16026.14	155	58.98
LCA	8	359	-6984.22	14686.44	16222.43	153	43.12
LCA	9	404	-6949.68	14707.37	16435.89	171	34.54
LCA	10	449	-6912.33	14722.66	16643.72	207	37.35
Rasch	1	87	-7763.41	15700.81	16073.04	70	
Rasch	2	173	-7405.26	15156.52	15896.71	250	
Rasch	3	259	-7217.40	14952.80	16060.94	250	
Mixed: Rasch/LCA	2	132	-7579.81	15423.63	15988.39	108	
Mixed: Rasch/LCA/LCA	3	177	-7342.05	15038.10	15795.40	250	
Mixed: Rasch/Rasch/LCA	3	218	-7279.43	14994.85	15927.57	250	

parsimonious 89 parameter two-class LCA. The AIC (See Figure 10) supported a three- to six-class model. The BIC (See Figure 10) and component log-likelihood (See Figure 11) values showed that a six-class solution fits best. This model was used in subsequent analysis. Table 7 displays the frequency of responses while the LCAs' relative class sizes are found in Table 8. The distribution of score estimates is in found in Figure 12. The mean score was a -0.16 with a standard deviation of 0.79. The raw score reliability was $\alpha \geq 0.80$ (Cronbach, 1951). None of the items in this section showed severe misfit (See Table 9).

Item Profiles

The item profiles for this section are complex (See Figures 13-17). Class 1 primarily chose category 1, meaning that this class believes that their evaluations were put to little use. Class 2 had proximity to both categories 2 and 3. This class believes that their evaluations were put to moderate use. Class 3 had a tendency to choose category 4. This is the class of subjects whose evaluations were put to the most extensive use. Class 4 often chose category 0, but had a pretty good proximity to the other options as well. These evaluations were put to primarily one or two uses. Class 5 had good proximity to response option 4 but also had a slight tendency toward category 3. This class believes that their evaluations had a high degree of use but not as to the degree of class 3. Class 6 primarily chose category 0. These subjects believe that their evaluations were not put to any use at all. The profiles for the different classes cross frequently and at all levels of the

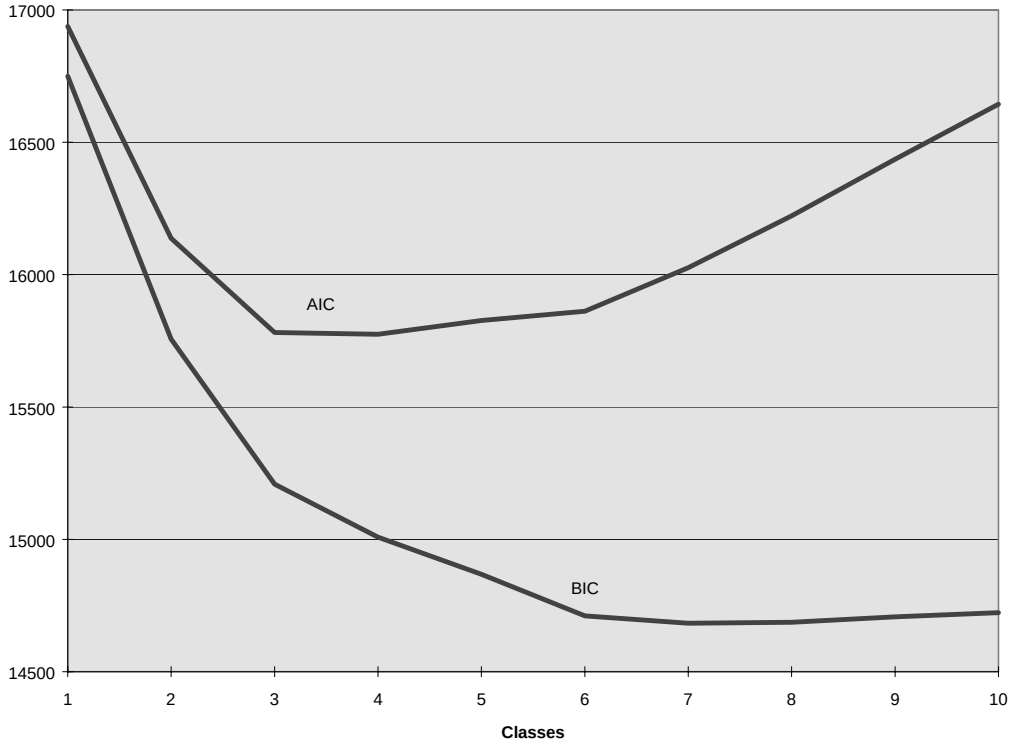

Figure 10. Information criteria for Utility of the Evaluation.

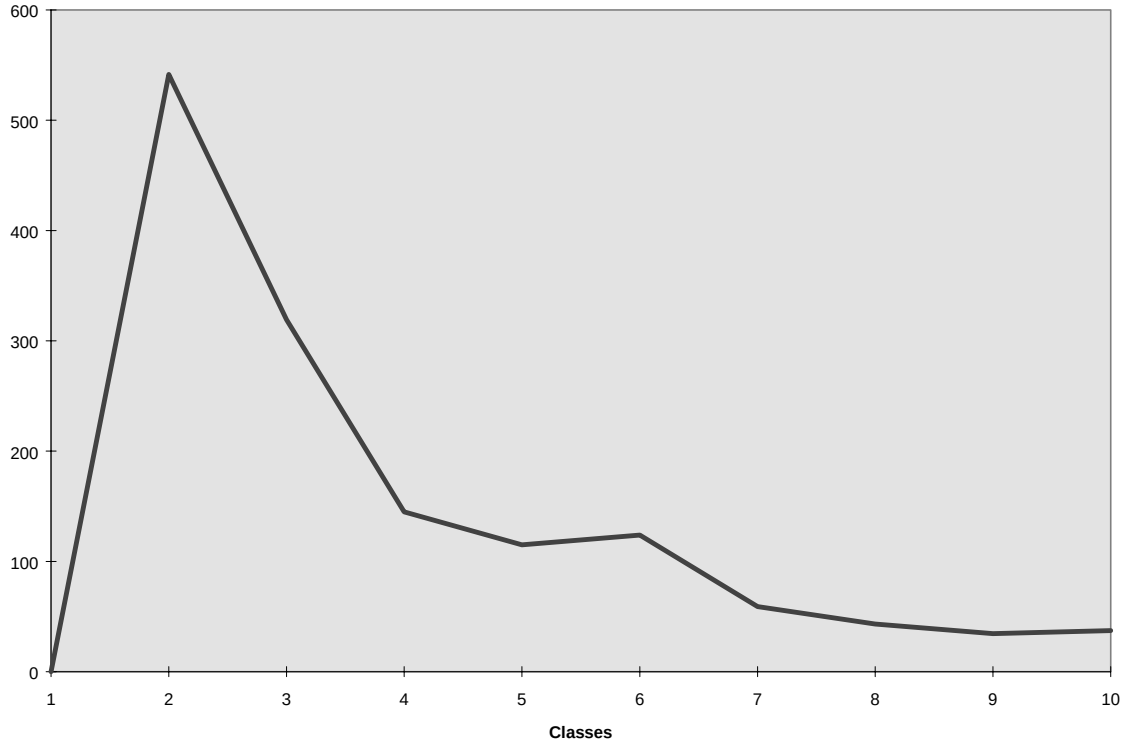

Figure 11. Component log-likelihood values for Utility of the Evaluation.

Table 7

Frequency of Subjects Using Each Response Option on Utility of the Evaluation

Item	Response					
	0	1	2	3	4	Missing
1	38	48	107	185	151	4
2	117	159	127	82	41	7
3	47	149	129	111	87	10
4	37	36	89	169	196	6
5	53	91	153	149	81	6
6	122	150	135	91	27	8
7	110	146	150	94	25	8
8	26	35	102	165	197	8
9	84	142	178	95	24	7
10	433	44	23	17	9	7
11	169	103	96	82	82	1

Table 8

Relative Class Size for Utility of the Evaluation

LCA 1	*LCA 2*	*LCA 3*	*LCA 4*	*LCA 5*	*LCA 6*	*LCA 7*	*LCA 8*	*LCA 9*	*LCA 10*
0.51832	0.48168								
0.43200	0.39131	0.17669							
0.41373	0.26766	0.16719	0.15142						
0.27987	0.22779	0.19888	0.15204	0.14142					
0.23346	0.22656	0.19739	0.16620	0.11816	0.05822				
0.21047	0.19108	0.17306	0.16971	0.14052	0.05831	0.05685			
0.20281	0.17306	0.16044	0.14199	0.09789	0.08804	0.07411	0.06168		
0.19488	0.18641	0.15974	0.11183	0.10788	0.10286	0.05830	0.05255	0.02555	
0.14891	0.13933	0.12781	0.11415	0.09484	0.08719	0.07598	0.07240	0.07171	0.06767

Rasch 1	*LCA 2*	*LCA 3*
0.54926	0.45074	
0.26235	0.38469	0.35295

Rasch 1	*Rasch 2*	*LCA 3*
0.74885	0.25115	
0.51026	0.19010	0.29964

Rasch 1	*Rasch 2*	*Rasch 3*
0.46049	0.35258	0.18693

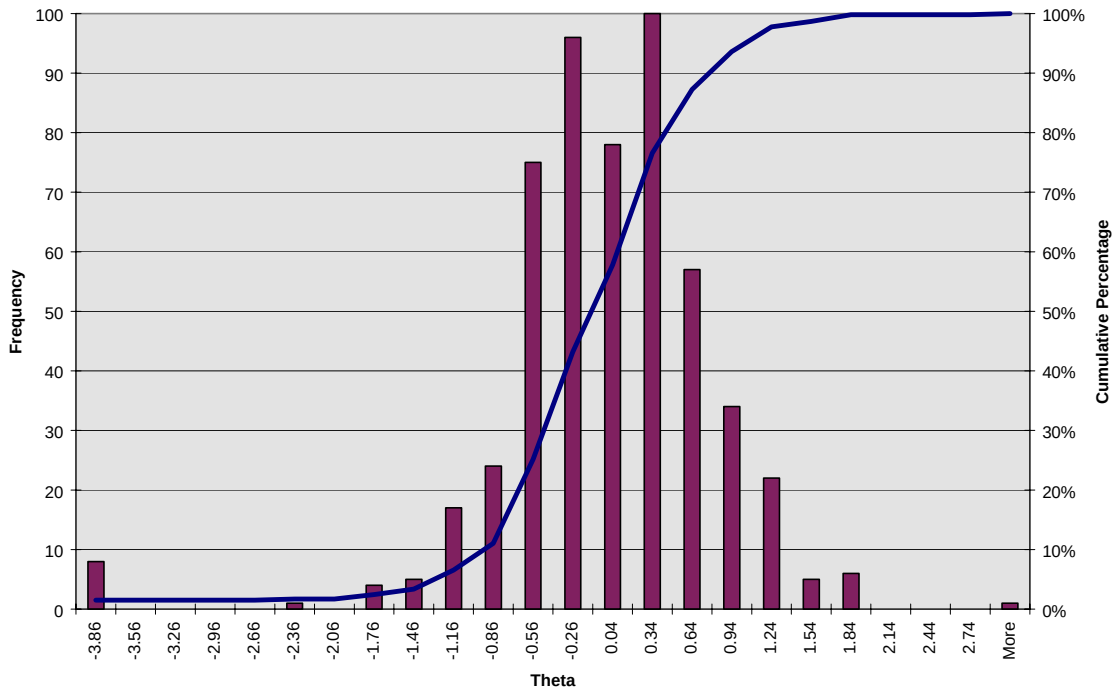

Figure 12. Frequency distribution of thetas for the ordinary Rasch model on Utility of the Evaluation.

Table 9

Item-Q-Index for Utility of the Evaluation

Item #	Q
1	0.1628
2	0.1378
3	0.2176
4	0.1469
5	0.1192
6	0.1060
7	0.0986
8	0.1581
9	0.1788
10	0.2775
11	0.2446

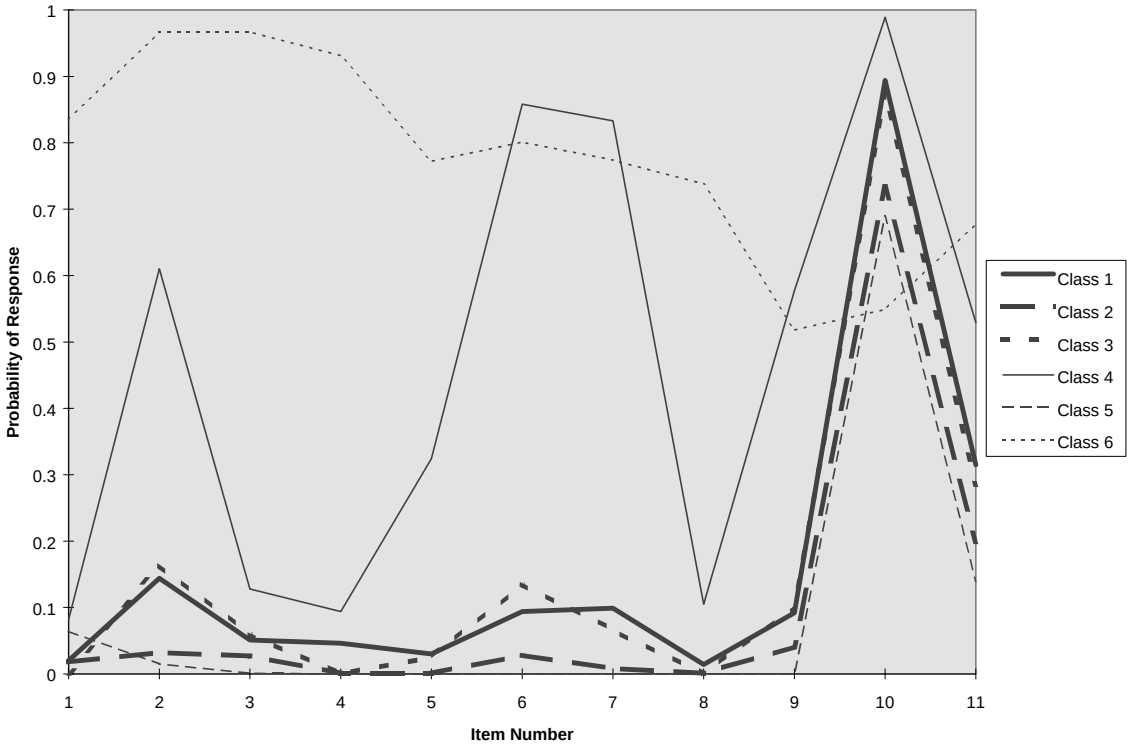

Figure 13. Each class's expected probability of responding with category 0 (Never Used) on Utility of the Evaluation.

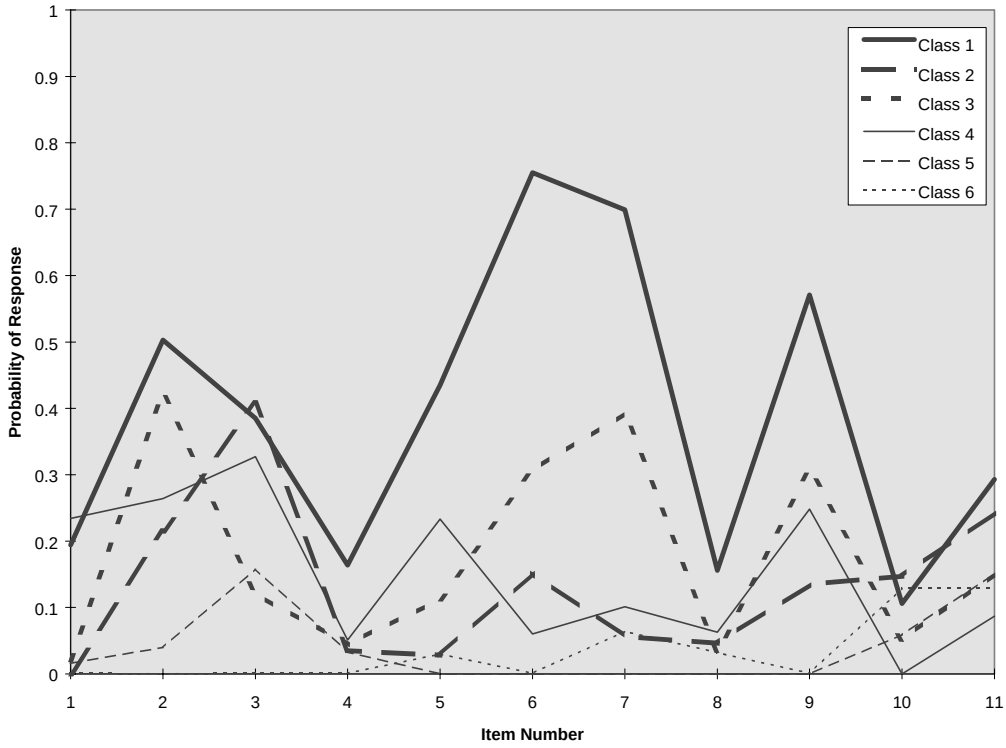

Figure 14. Each class's expected probability of responding with category 1 (Rarely Used) on Utility of the Evaluation.

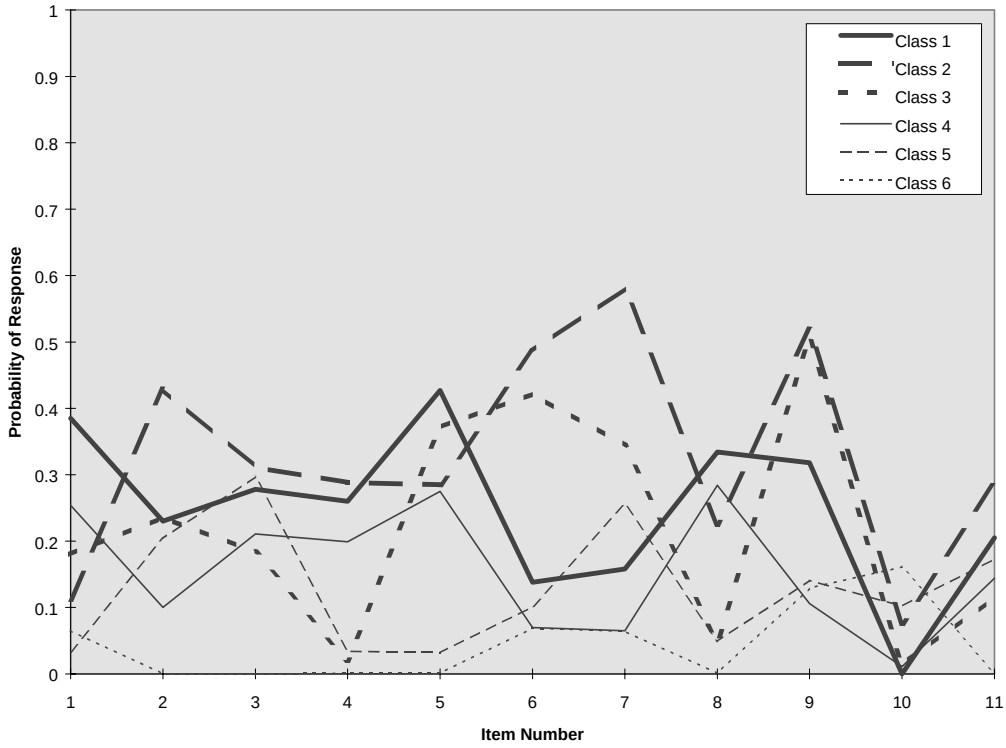

Figure 15. Each class's expected probability of responding with category 2 (Moderately Used) on Utility of the Evaluation.

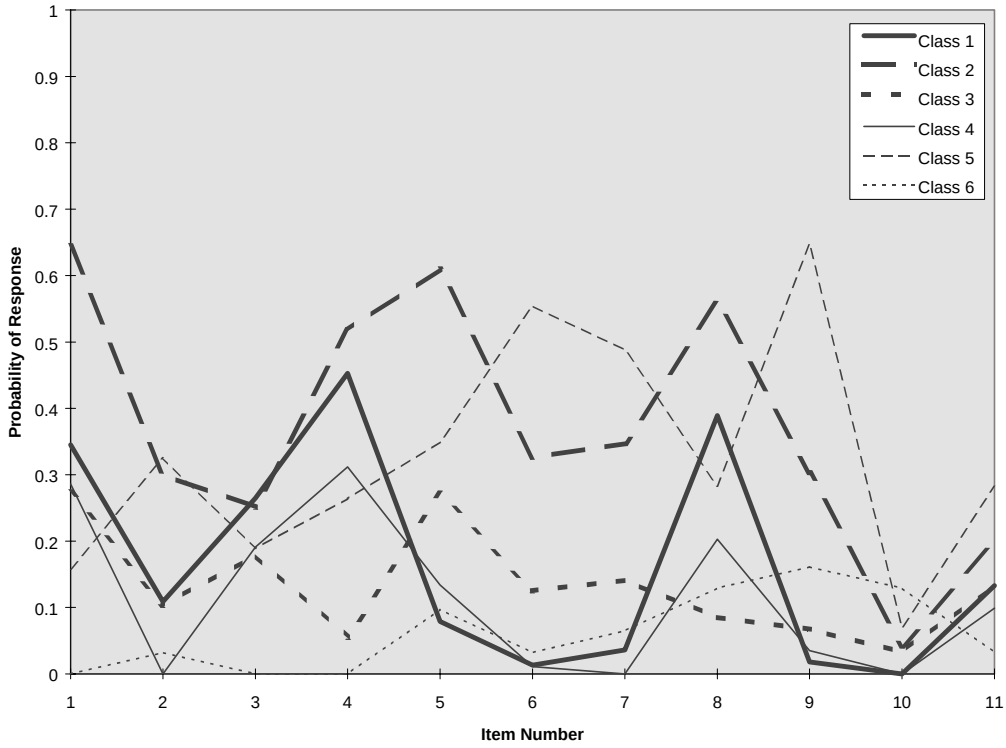

Figure 16. Each class's expected probability of responding with category 3 (Frequently Used) on Utility of the Evaluation.

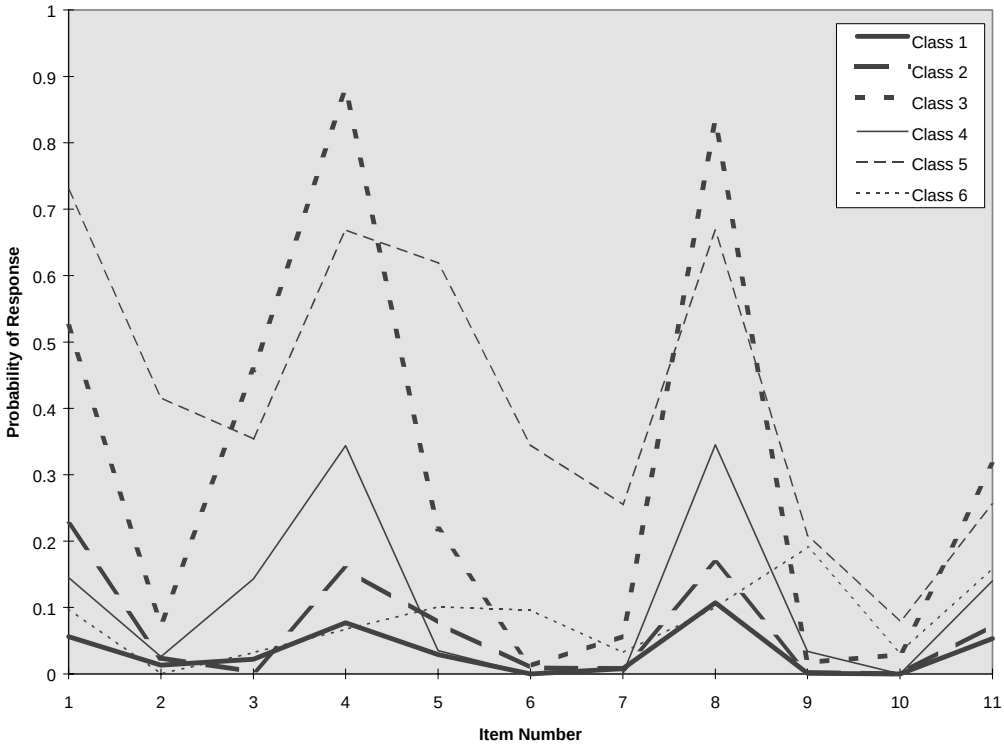

Figure 17. Each class's expected probability of responding with category 4 (Extensively Used) on Utility of the Evaluation.

manifest variables. This suggests that multi-dimensionality may be occurring for this section.

Distribution of Latent Classifications

The frequency of classification at each level of the Rasch latent continuum adds further testimony to this claim (See Figure 18). Notice how the distributions of Class 2 and Class 5 seem to be on top of each other. This is also true of Classes 1 and 4.

Discussion

In all, the item profiles and the frequency of classifications by score indicate that dimensionality and the Rasch model may not be appropriate for this section. This is contrary to the goodness of fit information, which revealed that the data fit the Rasch model better than the comparable two-class LCA. Since the item profiles and the distribution of classifications were derived from the less than parsimonious six-class LCA model, the ordinary Rasch model is probably the best model for this section. Future investigation into more parsimonious LCA models might be useful to decipher this section more.

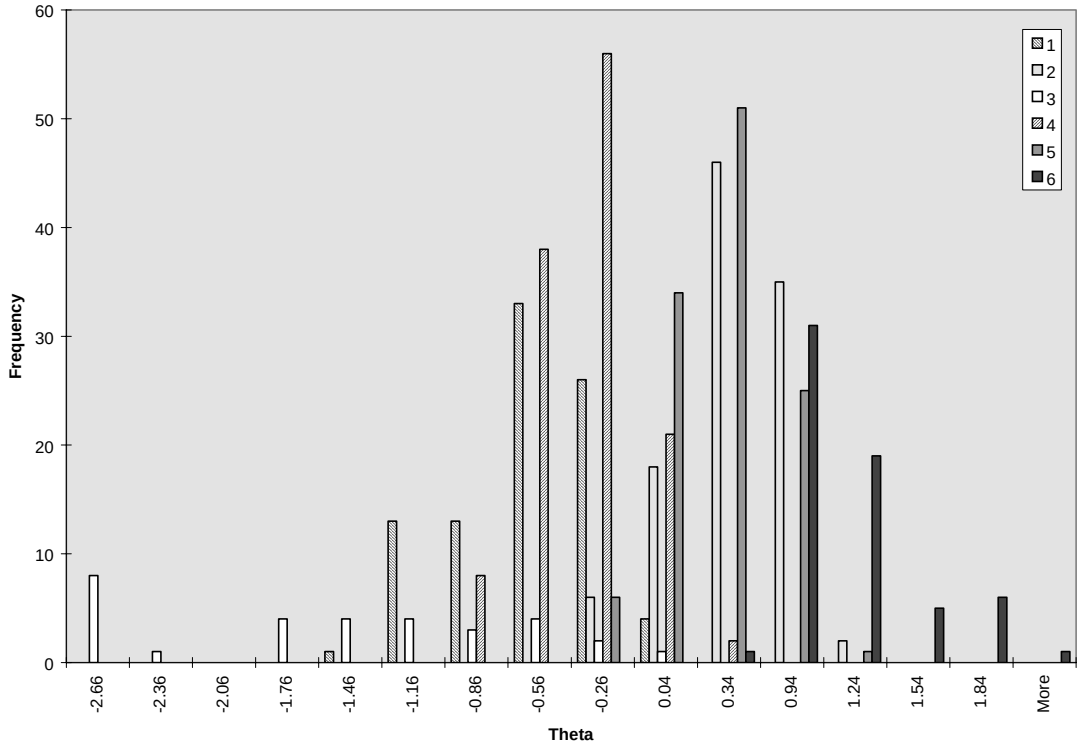

Figure 18. Distribution of subjects within each classification of the 6 Class LCA at various levels of estimated scores from the ordinary Rasch model on Utility of the Evaluation.

Section 3: Attitudes Toward Evaluation

Goodness of Fit

Goodness of fit statistics for this section can be found in Table 10. The data did not fit the Rasch model as well as the previous two sections. Only the one-class LCA fit worse. Based on the AIC (See Figure 19) a three-, four-, or five- class LCA fit best. The BIC and the log-likelihood values (See Figure 20) provided no additional information. A three class solution was deemed best fitting due to its simplicity. The frequency of responses are found in Table 11 while the relative LCAs' class sizes are in Table 12. The mean of score estimates for this section was -0.46 with a standard deviation of 1.07. The raw score reliability was $\alpha \geq 0.85$ (Cronbach, 1951). The distribution of score estimates can be found in Figure 21. There were a great number of items that misfit this section (See Table 13). Items 10, 11, 13, 14, 15, 16, 18, 19, 21, 22, and 25 all show signs of severe misfit.

Item Profiles

The item profiles for this section are most interesting due to availability of the undecided response option (See Figures 22-26). The proximity of any class to this response was very low. Class 1 had a high tendency toward response 1 and 2. This class consists of those individuals without a strong conviction toward any of the comments. They neither strongly agree nor strongly disagree. Class 2 often chose response 3 for the

Table 10

Goodness of Fit Statistics for Attitudes Toward Evaluation

Model	*Classes*	*Parameters*	*Log-Likelihood*	*AIC*	*BIC*	*Iterations*	*Component Log-Likelihood*
LCA	1	100	-17299.47	34798.94	35225.47	44	0
LCA	2	201	-16436.73	33275.45	34132.78	63	862.74
LCA	3	302	-15842.68	32289.35	33577.47	142	594.05
LCA	4	403	-15496.63	31799.26	33518.18	223	346.05
LCA	5	504	-15236.67	31481.33	33631.04	208	259.96
LCA	6	605	-15045.80	31301.61	33882.11	201	190.87
LCA	7	706	-14868.21	31148.41	34159.71	250	177.59
LCA	8	807	-14720.04	31054.08	34496.17	245	148.17
LCA	9	908	-14565.55	30947.10	34819.99	250	154.49
LCA	10	1009	-14500.25	31018.50	35322.19	250	65.30
Rasch	1	199	-16900.47	34198.95	35047.74	122	
Rasch	2	397	-16009.52	32813.03	34506.36	113	
Rasch	3	595	-15427.19	32044.38	34582.23	250	
Mixed: Rasch/LCA	2	300	-16180.41	32960.83	34240.42	250	
Mixed: Rasch/LCA/LCA	3	401	-15654.35	32110.70	33821.09	250	
Mixed: Rasch/Rasch/LCA	3	498	-15527.97	32051.94	34176.06	88	

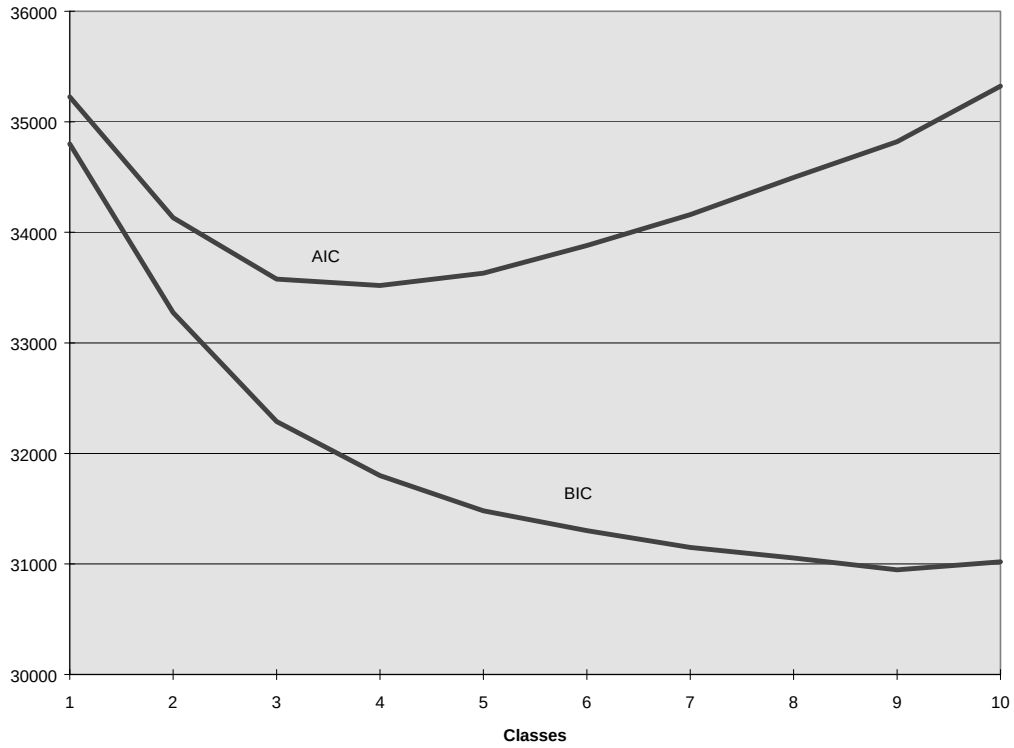

Figure 19. Information criteria for Attitudes Toward Evaluation.

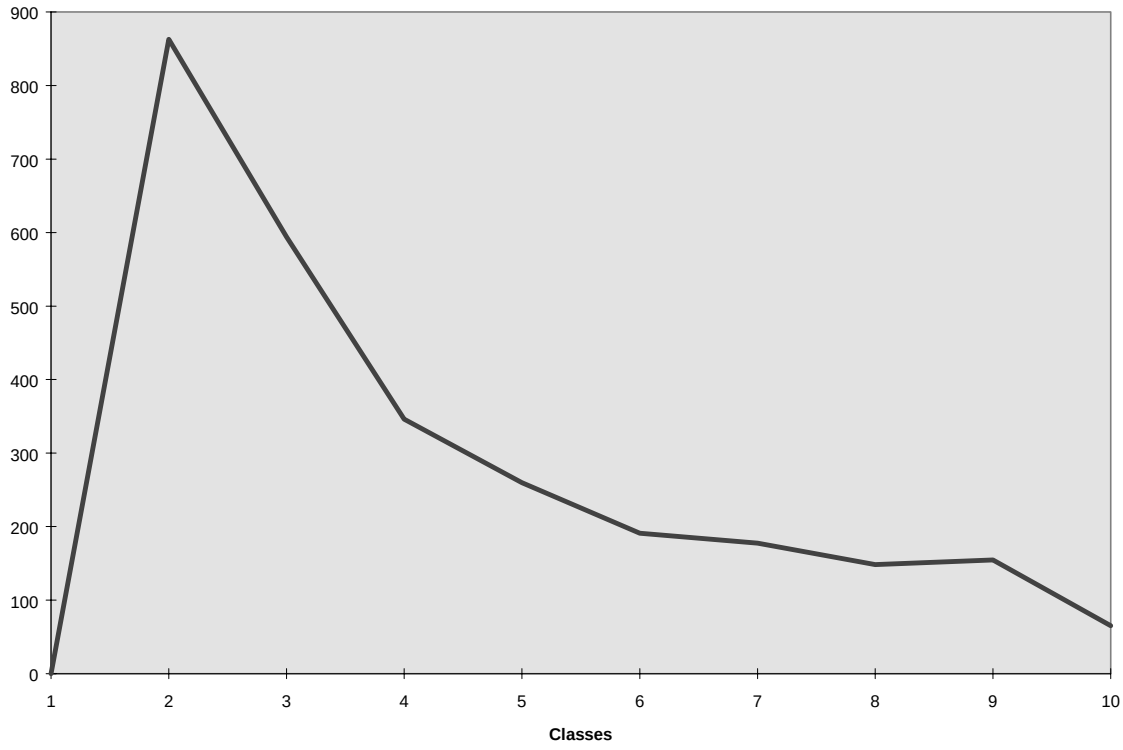

Figure 20. Component log-likelihood for Attitudes Toward Evaluation.

Table 11

Frequency of Subjects Using Each Response Option on Attitudes Toward Evaluation

Item	Response					
	0	1	2	3	4	Missing
1	129	131	141	121	8	3
2	80	117	228	80	27	1
3	39	6	140	339	5	4
4	70	136	208	83	32	4
5	56	74	280	100	18	5
6	84	198	175	66	8	2
7	49	79	287	93	22	3
8	142	148	138	55	47	3
9	56	93	252	118	12	2
10	39	47	271	139	33	4
11	48	62	234	176	9	4
12	99	184	120	75	48	7
13	41	58	295	110	25	4
14	121	184	158	54	11	5
15	105	197	136	86	5	4
16	127	296	81	19	6	4
17	47	116	278	58	29	5
18	30	117	202	137	40	7
19	35	208	188	63	34	5
20	83	320	80	26	23	1
21	60	181	195	71	24	2
22	220	228	55	9	16	5
23	73	163	185	61	40	11
24	66	219	137	46	54	11
25	11	29	226	250	15	2

Table 12

Relative Class Size for Attitudes Towards Evaluation

LCA 1	*LCA 2*	*LCA 3*	*LCA 4*	*LCA 5*	*LCA 6*	*LCA 7*	*LCA 8*	*LCA 9*	*LCA 10*
0.68448	0.31552								
0.46312	0.30624	0.23064							
0.34562	0.27681	0.24844	0.12914						
0.35828	0.25273	0.19156	0.13472	0.06272					
0.27913	0.22742	0.16422	0.13930	0.12534	0.06459				
0.19584	0.17800	0.15636	0.15319	0.13447	0.10495	0.07719			
0.18373	0.17193	0.14440	0.14078	0.13397	0.11234	0.05962	0.05322		
0.17112	0.14253	0.13385	0.12699	0.11444	0.11132	0.10357	0.05892	0.03726	
0.15211	0.14835	0.13699	0.11485	0.09580	0.09384	0.09319	0.08512	0.05513	0.02463

Rasch 1	*LCA 2*	*LCA 3*
0.48772	0.51228	
0.31487	0.45133	0.23379

Rasch 1	*Rasch 2*	*LCA 3*
0.56616	0.43384	
0.31030	0.29390	0.39580

Rasch 1	*Rasch 2*	*Rasch 3*
0.40026	0.30369	0.29604

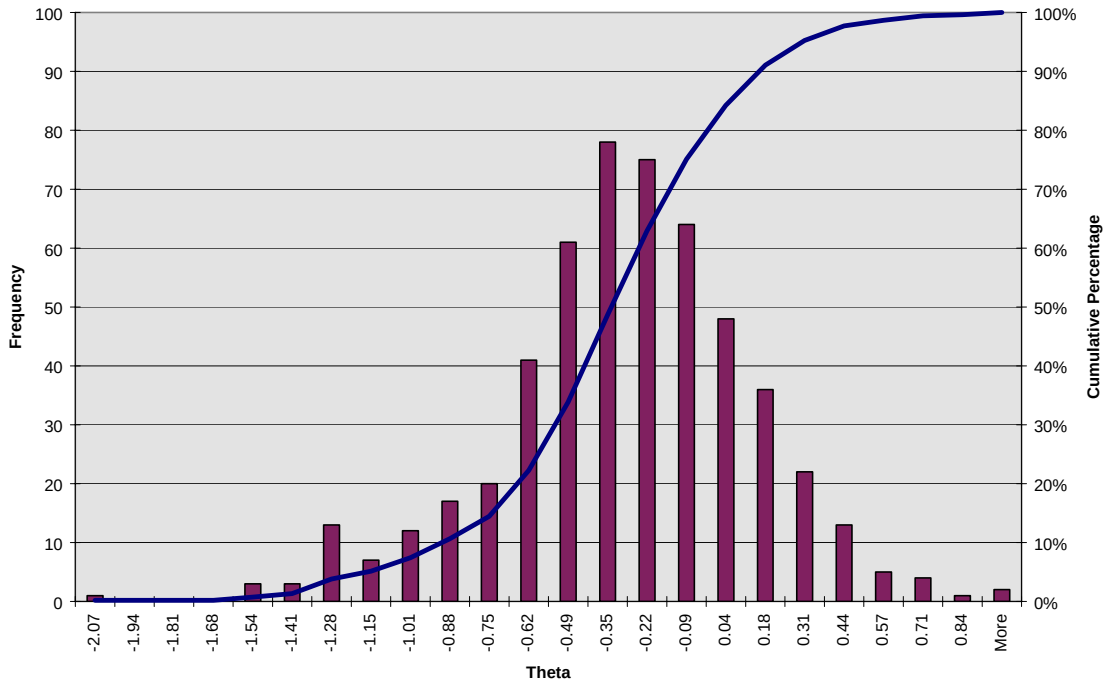

Figure 21. Frequency distribution of thetas for the ordinary Rasch model on Attitudes Toward Evaluation.

Table 13

Item-Q-Index for Attitudes Toward Evaluation

Item #	Q
1	0.2350
2	0.2359
3	0.2440
4	0.2357
5	0.2059
6	0.2599
7	0.2271
8	0.2243
9	0.2906
10	0.3656
11	0.3170
12	0.2538
13	0.3197
14	0.3581
15	0.3881
16	0.4092
17	0.2934
18	0.3154
19	0.3796
20	0.2876
21	0.3164
22	0.3484
23	0.2447
24	0.2801
25	0.4122

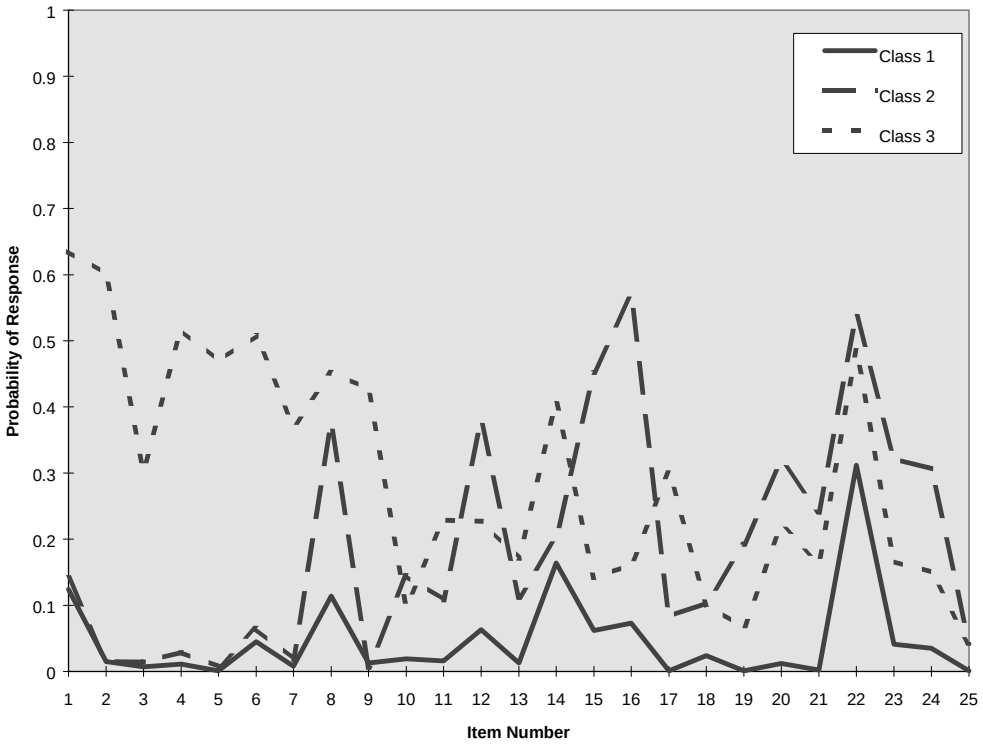

Figure 22. Each class's expected probability of responding with category 0 (Strongly Disagree) on Attitudes Toward Evaluation.

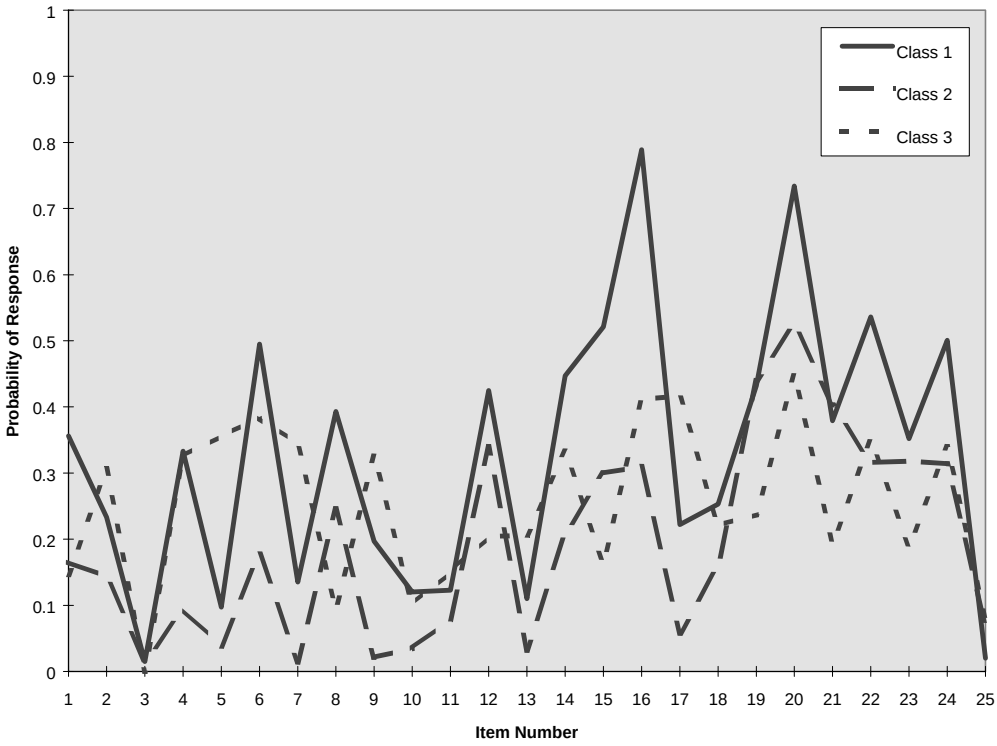

Figure 23. Each class's expected probability of responding with category 1 (Disagree) on Attitudes Toward Evaluation.

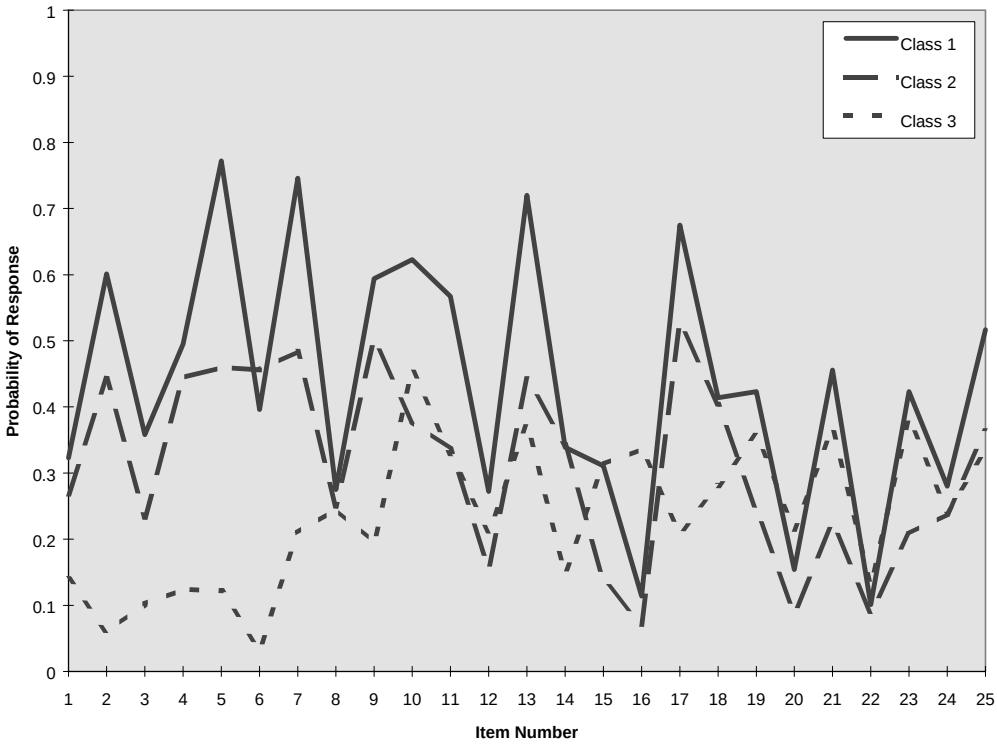

Figure 24. Each class's expected probability of responding with category 2 (Agree) on Attitudes Toward Evaluation.

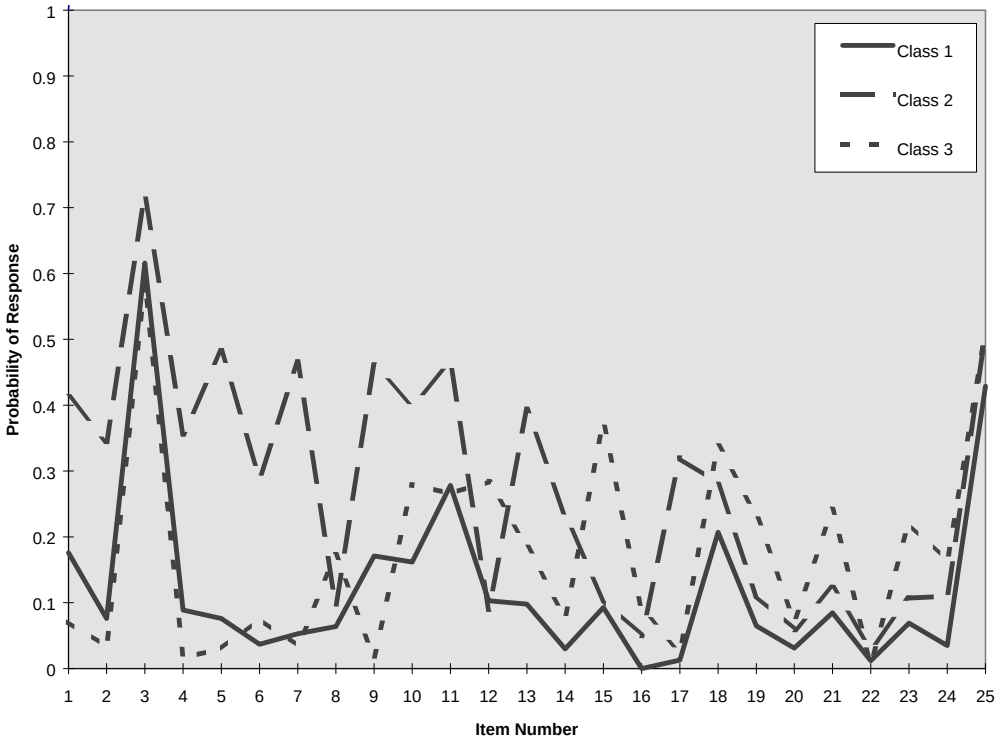

Figure 25. Each class's expected probability of responding with category 3 (Strongly Agree) on Attitudes Toward Evaluation.

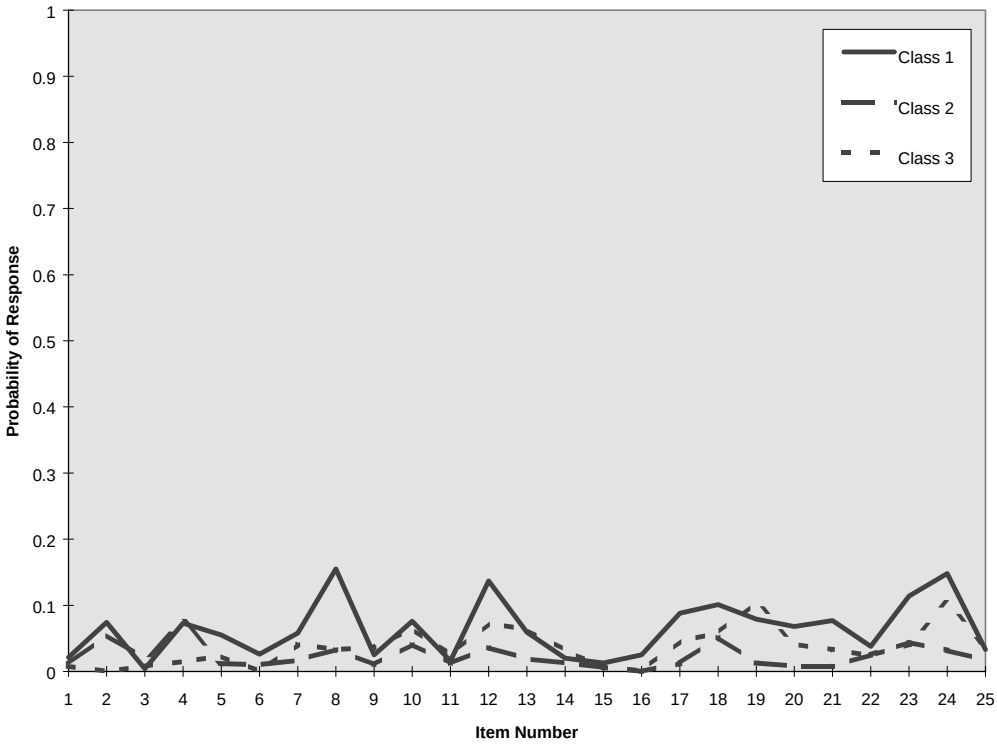

Figure 26. Each class's expected probability of responding with category 4 (Undecided) on Attitudes Toward Evaluation.

first half of the section and response 0 for the second half. They tend to agree with many of the statements in the first half of the section and disagree with those in the second half. On the other hand, Class 3 behaves just the opposite. These subjects primarily chose response 0 in the first half and response 3 in the second half. It should be noted that the profile for the second half of the section, which contained the misfittting items, was still capable of displaying disparity between the two extreme classes. Further investigation into the misfitting survey questions might reveal that there are at least two dimensions along which these types of statements pertain. Class 2 might agree with the first type of statement and disagree with the other. While Class 3 might disagree with the first type and agree with the second. In all, there seems to be three distinct classes: those who have strong opinions one way, those who have strong opinions the opposite way, and those who ride the fence.

Distribution of Latent Classifications

The distribution of Rasch score estimates for Class 1 are expectedly well distributed (See Figure 27). However, the scores from Class 3 are definitely higher than those of the other two classes while the scores for Class 2 are definitely lower than the other two classes.

Discussion

Since there is such a high number of misfitting items, content loss would be an issue if these items were deleted in order to improve the fit of the data to any of the models. As

a

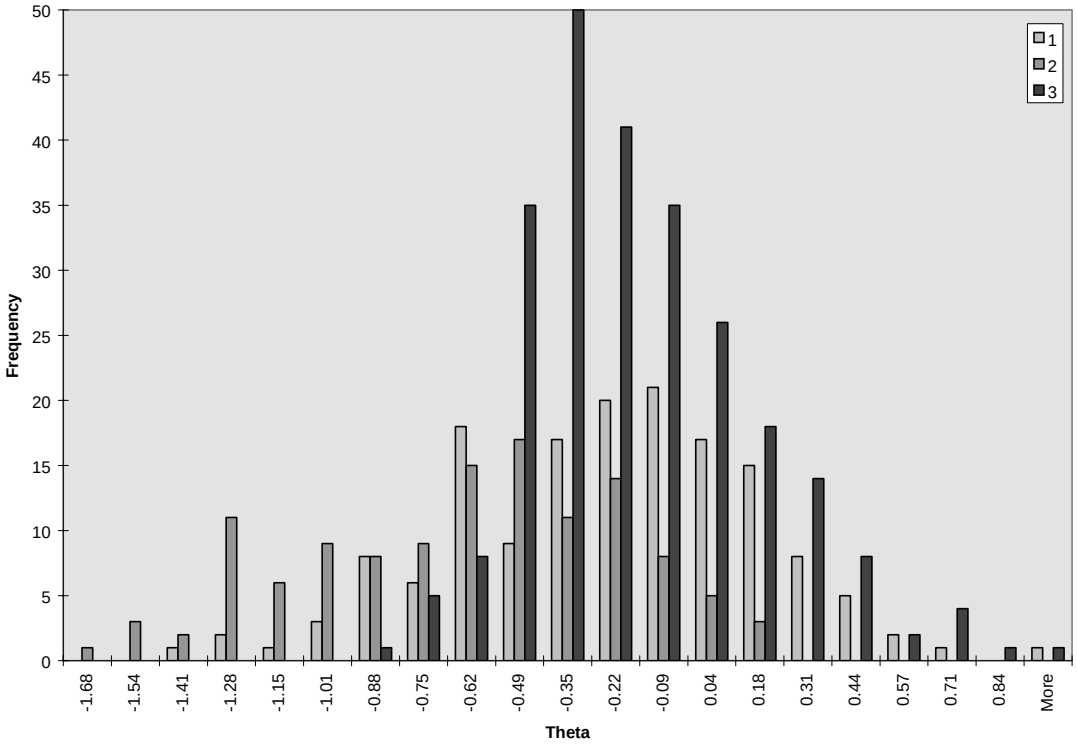

Figure 27. Distribution of subjects within each classification of the 3 Class LCA at various levels of estimated scores from the ordinary Rasch model on Attitudes Toward Evaluation.

result, this three-class unordered LCA is the best fitting model and provides good information using all of the questions in the section including those which misfit the Rasch model.

Section 4: Combination of All Three Sections

Goodness of Fit

The goodness of fit statistics revealed that the six-class LCA fit this section best (See Table 14). The data did not fit the ordinary Rasch model as well as they fit any of the LCAs, except for the one-class LCA. The BIC flattened out at six classes while the component log-likelihood was modestly high for this model also (See Figures 28 and 29). Although the AIC could have supported a more parsimonious two- or four- class model, the information provided by the BIC and component log-likelihood deemed the six-class LCA the best fit. Relative class size can be seen in Table 15. For this section, the MIRA and hybrid models were too complex for the computer to handle. After repeated computer lock ups, these models were not attempted again. The mean score estimate was -0.14 with a standard deviation of 1.09. The raw score reliability was $\alpha \geq 0.90$ (Cronbach, 1951). The distribution of score estimates can be found in Figure 30. Items 3, 10, and 11 from the Utility section and all items after item 8 in the Attitudes section showed severe misfit (See Table 16).

Table 14

Goodness of Fit Statistics for All Three Sections Combined

Model	*Classes*	*Parameters*	*Log-Likelihood*	*AIC*	*BIC*	*Iterations*	*Component Log-Likelihood*
LCA	1	204	-35891.20	72190.4	73060.52	43	0
LCA	2	409	-34381.72	69581.43	71325.94	70	1509.48
LCA	3	614	-33427.64	68083.29	70702.18	138	954.08
LCA	4	819	-32887.71	67413.42	70906.71	242	539.93
LCA	5	1024	-32567.39	67182.78	71550.45	192	320.32
LCA	6	1229	-32128.51	66715.02	71957.08	250	438.88
LCA	7	1434	-31876.95	66621.90	72738.34	250	251.56
LCA	8	1639	-31616.16	66510.32	73501.15	248	260.79
LCA	9	1844	-31376.72	66441.44	74306.65	250	239.44
LCA	10	2049	-31171.79	66441.58	75181.18	250	204.93
Rasch	1	407	-34853.15	70520.31	72256.28	250	

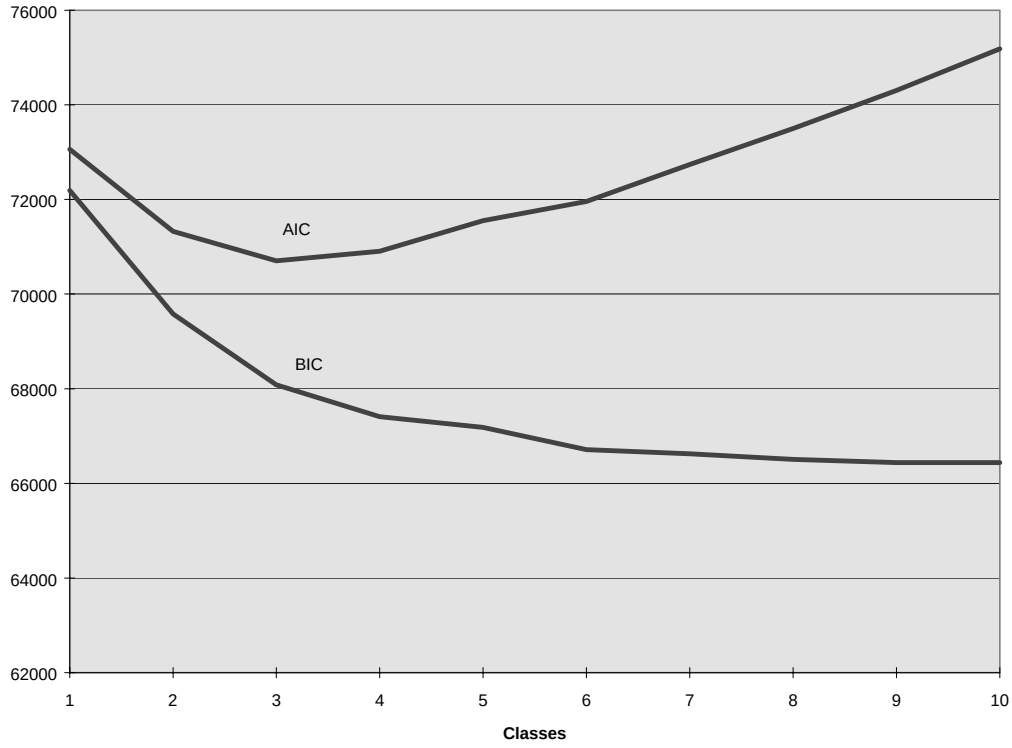

Figure 28. Information criteria for all three sections combined.

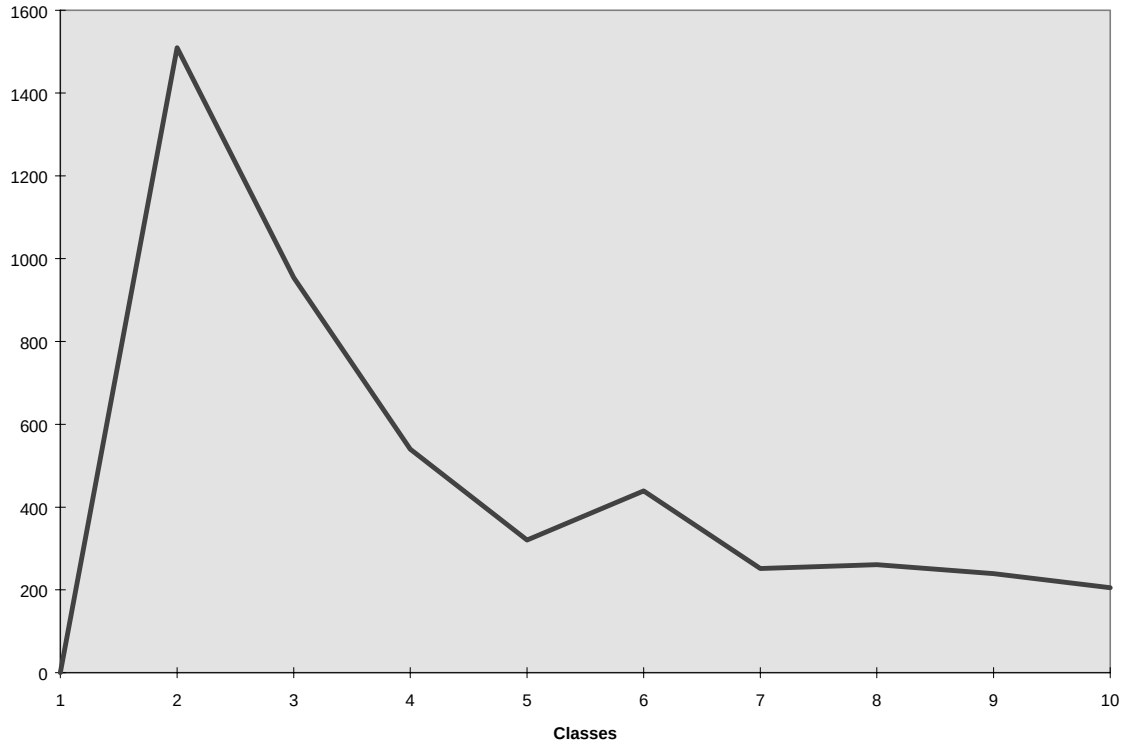

Figure 29. Component log-likelihood for all three sections combined.

Table 15

Relative Class Size for All Three Sections Combined

LCA 1	*LCA 2*	*LCA 3*	*LCA 4*	*LCA 5*	*LCA 6*	*LCA 7*	*LCA 8*	*LCA 9*	*LCA 10*
0.51206	0.48794								
0.40227	0.33733	0.26040							
0.37640	0.21968	0.20460	0.19931						
0.25905	0.21264	0.20633	0.17306	0.14893					
0.22876	0.18495	0.16618	0.15002	0.14053	0.12955				
0.21873	0.19180	0.14128	0.13427	0.12738	0.09912	0.08742			
0.18173	0.17755	0.15887	0.12716	0.12259	0.11053	0.07218	0.04939		
0.15205	0.13987	0.12764	0.12241	0.11696	0.11418	0.10716	0.07219	0.04755	
0.20831	0.14440	0.13380	0.12910	0.11040	0.08395	0.06080	0.04943	0.04940	0.03042

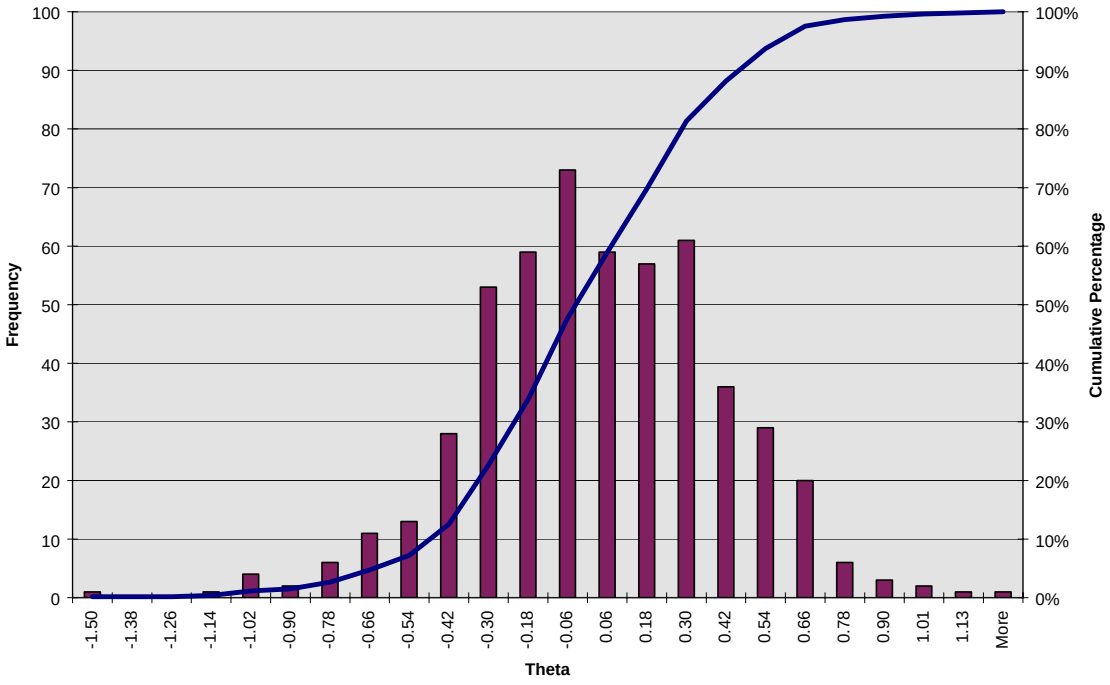

Figure 30. Frequency distribution of thetas for the ordinary Rasch model on all three sections combined.

Table 16

Item-Q-Index for All Three Sections Combined

Item #	Q	Item #	Q	Item #	Q
Content 1	0.2555	Utility 1	0.2543	Attitude 1	0.2705
Content 2	0.2223	Utility 2	0.2258	Attitude 2	0.2711
Content 3	0.2448	Utility 3	0.3512	Attitude 3	0.3362
Content 4	0.2376	Utility 4	0.2773	Attitude 4	0.2808
Content 5	0.2216	Utility 5	0.2389	Attitude 5	0.2679
Content 6	0.2092	Utility 6	0.2262	Attitude 6	0.2445
Content 7	0.1982	Utility 7	0.2168	Attitude 7	0.2925
Content 8	0.1997	Utility 8	0.2979	Attitude 8	0.3708
Content 9	0.2460	Utility 9	0.2198	Attitude 9	0.2552
Content 10	0.2571	Utility 10	0.3256	Attitude 10	0.4135
Content 11	0.2100	Utility 11	0.3255	Attitude 11	0.4085
Content 12	0.2479			Attitude 12	0.4211
Content 13	0.2060			Attitude 13	0.3943
Content 14	0.2612			Attitude 14	0.3615
Content 15	0.2558			Attitude 15	0.5404
				Attitude 16	0.5425
				Attitude 17	0.3258
				Attitude 18	0.4075
				Attitude 19	0.4767
				Attitude 20	0.4156
				Attitude 21	0.4327
				Attitude 22	0.4202
				Attitude 23	0.3946
				Attitude 24	0.4153
				Attitude 25	0.4922

Item Profiles

In this situation, where there are six classes, as well as 51 items broken down into three sections, item profile interpretation is a difficult task (See Figure 31-35). Nonetheless, Class 1 seems to be the class with little conviction. These subjects tended toward response categories 1, 2, and 3. They rarely used response option 0 or 4. On Content of the Evaluation, they tended more toward response 3 (Moderately Used) than response 1 (Rarely Used or Disagree); while on Utility and Attitudes they tended more toward response 1 than response 3.

Class 2 had a high probability of responding to category 3 (Frequently Used) for both content and utility. This class generally strongly agrees with the first half of the Attitudes statements and thus chose response 4 while maintaining good distribution over the second half of the section.

The content of the evaluations for Class 3 had a wide variety as well as good utility seen by these subjects' use of categories 3 and 4 (Frequently or Extensively Used). On the Attitudes section, this class primarily chose responses 2 and 3 meaning that these subjects have little conviction regarding any of the statements.

Class 4 had an overwhelming probability to use response 4 on content and utility This class maintains that their evaluations covered a broad area and were put to good use. This class tended to choose response options 0 and 3 on the Attitudes section, showing that they have strong convictions either positively or negatively for the Attitudes statements.

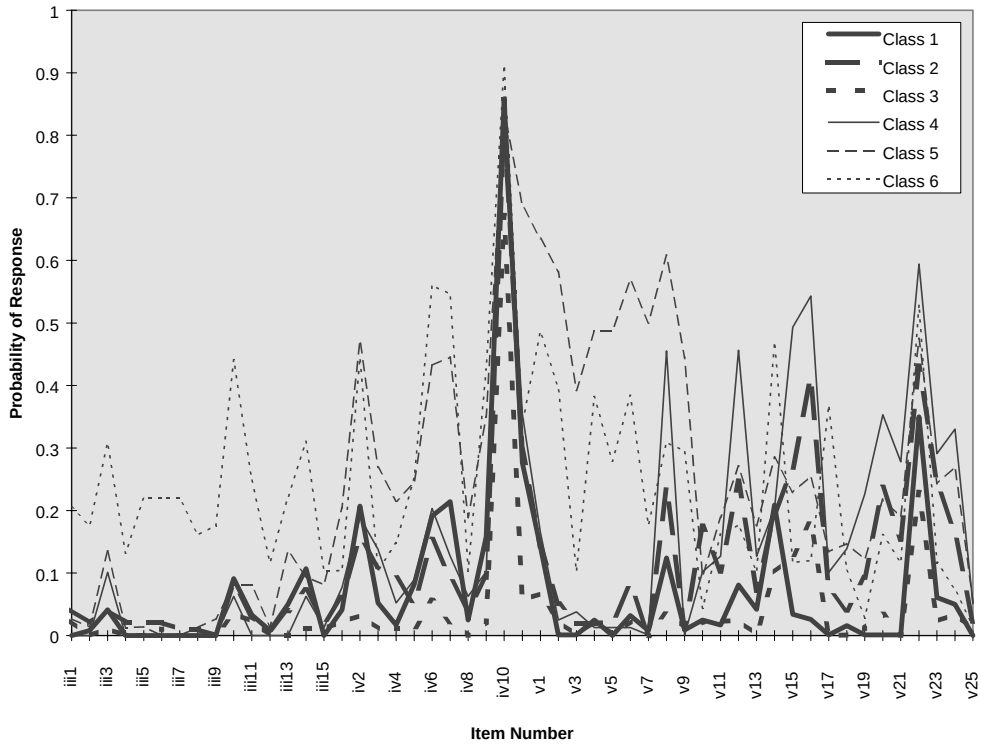

Figure 31. Each class's expected probability of responding with category 0 on all three sections combined.

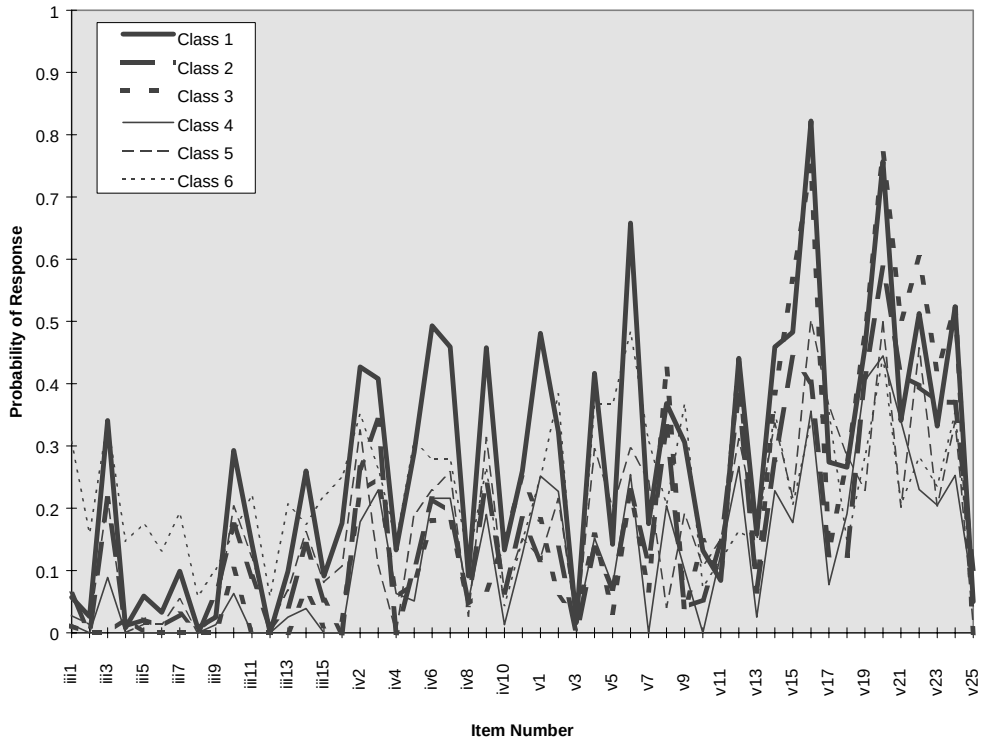

Figure 32. Each class's expected probability of responding with category 1 on all three sections combined.

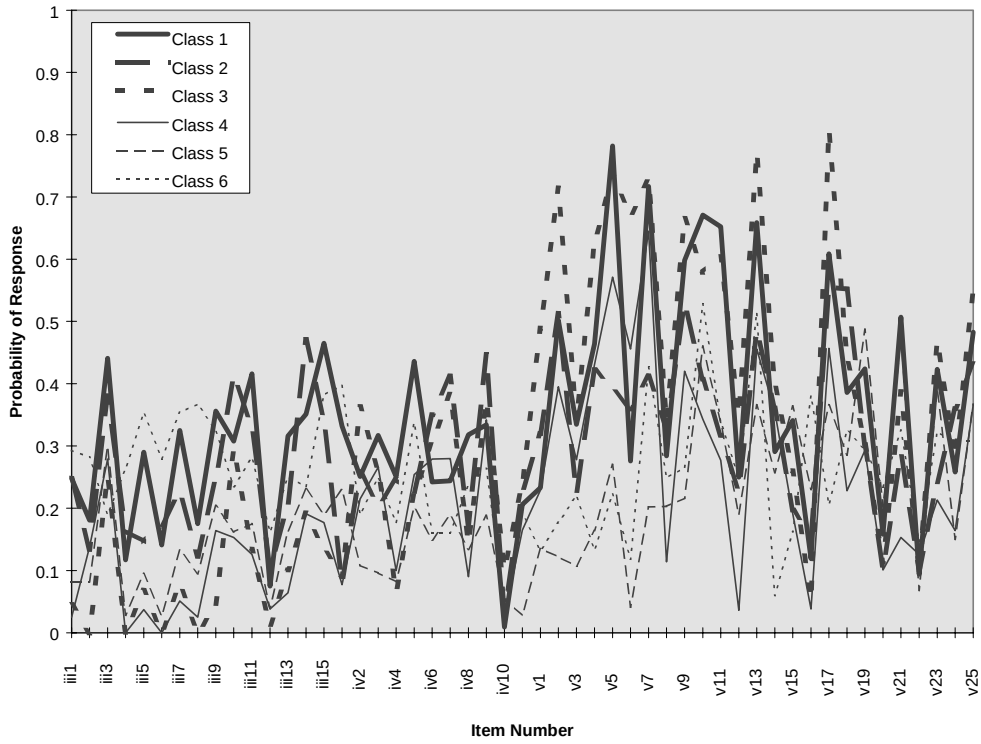

Figure 33. Each class's expected probability of responding with category 2 on all three sections combined.

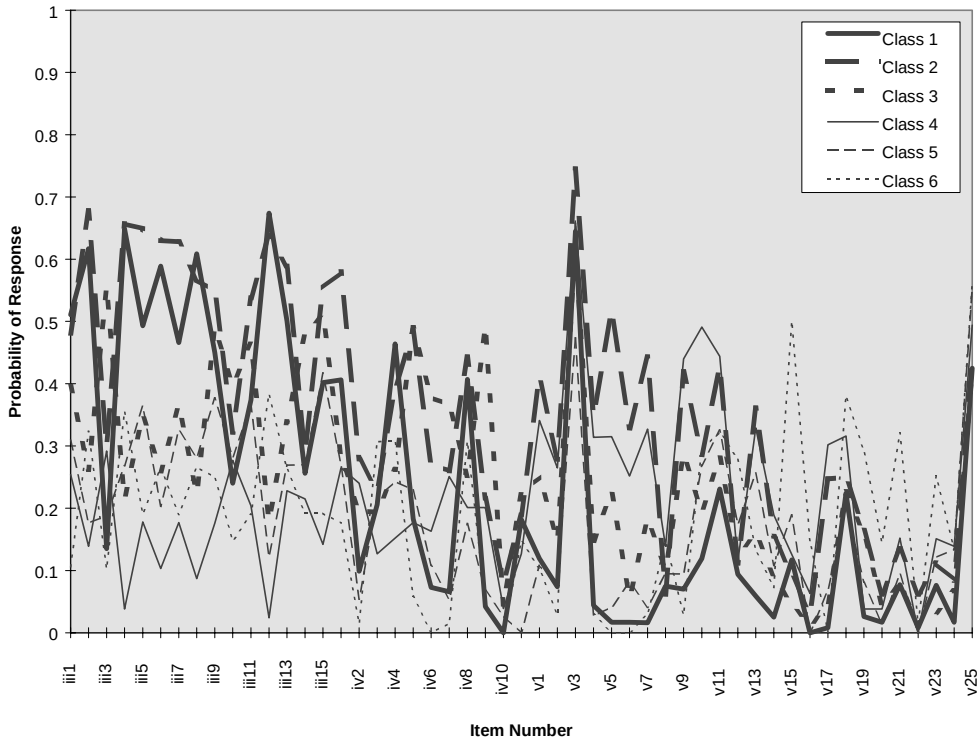

Figure 34. Each class's expected probability of responding with category 3 on all three sections combined.

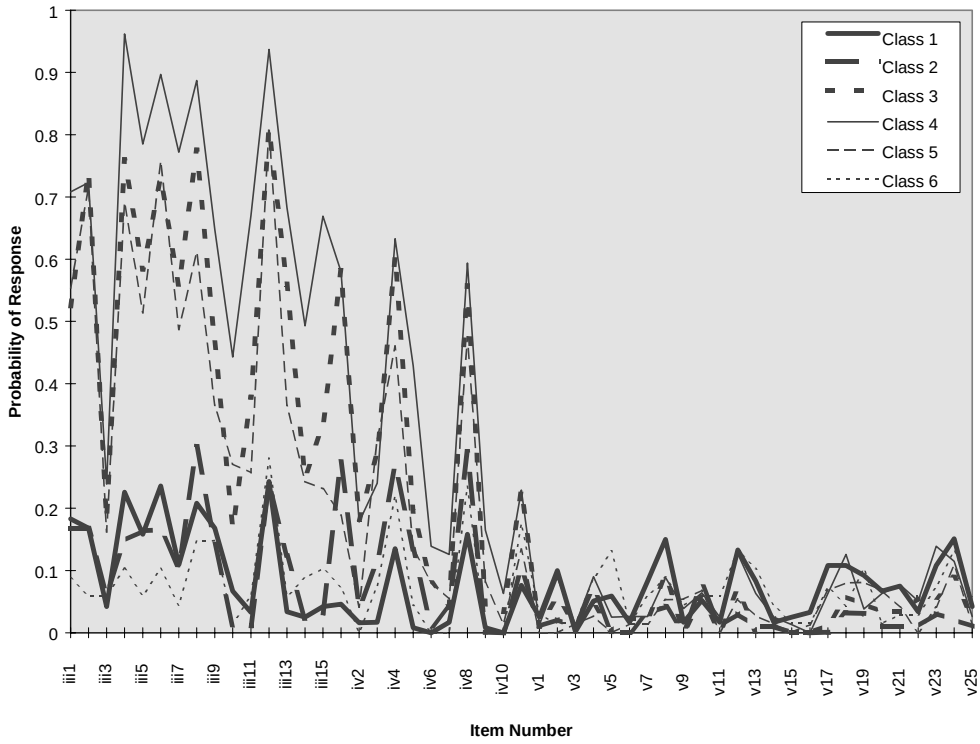

Figure 35. Each class's expected probability of responding with category 4 on all three sections combined.

In Content of the Evaluation, Class 5 chose category 4 often and category 0 rarely. These subjects' evaluations covered a wide area of content. However, on Utility they frequently chose response 0, displaying that these evaluations were put to little use. Still, on some of the various questions in this section they had a high probability towards response 4, indicating that these subjects' evaluations were used primarily for only one or two purposes. On the first half of the Attitudes section, this class often chose response 0, indicating that they strongly disagree with the statements Responses to the second half of the section were well distributed over all responses.

Class 6 generally believes that their evaluations covered little content and were put to little use by having responded primarily to categories 0 and 1 on these sections. On the Attitudes section, this class chose low responses on the first half and high responses on the second half, indicating that they strongly disagree with the first half and strongly agree with the second half.

Distribution of Classifications

Since the combination of all three sections represents at least three different traits a continuous level of measurement is nonsensical for use. Still, the frequency distribution of classifications reveals that these classes do have some order if they are broken down into three groups (See Figure 36). Classes 2 and 5 had low scores. Classes 1 and 6 had average scores while Classes 3 and 4 had high scores. There was some differentiation between scores within each of these three groups making ordinal classification possible.

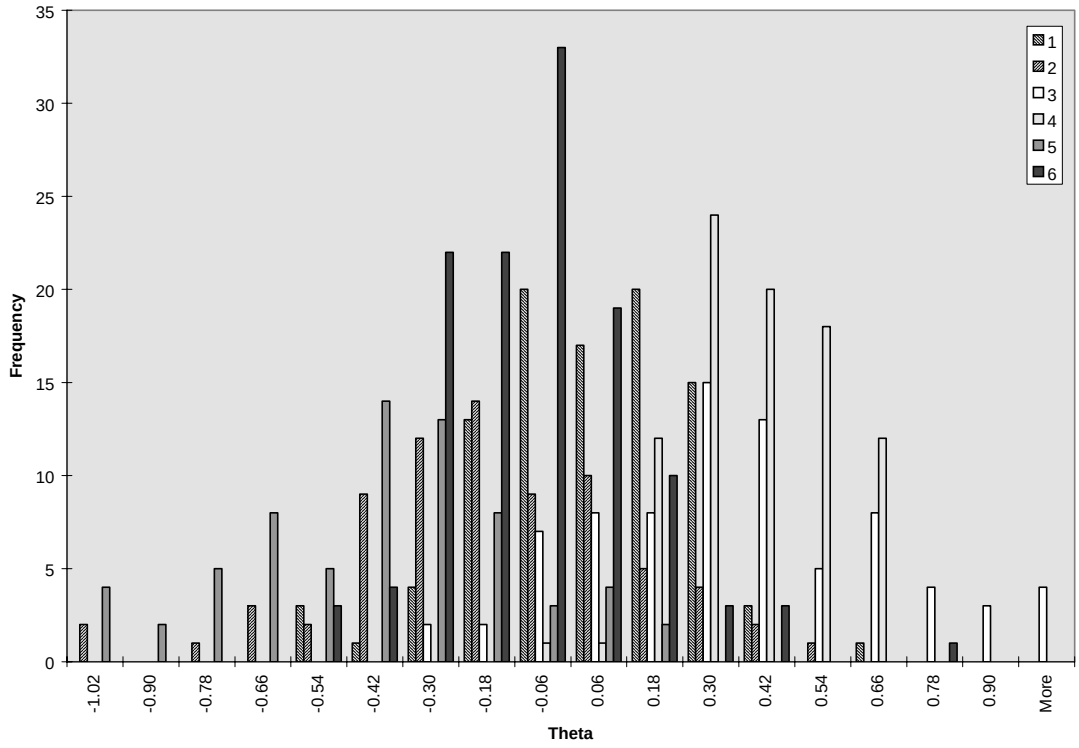

Figure 36. Distribution of subjects within each classification of the 6 Class LCA at various levels of estimated scores from the ordinary Rasch model on all three sections combined.

Discussion

Due to the extreme amount of overlap in distributions and the a priori knowledge of the multi-dimensionality of this combined section, an unordered six-class LCA is the most preferable model. The goodness of fit of the two comparable models, the two-class LCA and the Rasch model also supported this claim. Still, investigation into the three-class model as well as other more parsimonious models may reveal that ordered classes may be possible, but using the Rasch model with or without the misfitting items is not appropriate.

CONCLUSION

Implications

The various latent class models used in this study were very capable of modeling the data on the two sections of this questionnaire that included severely misfitting items. For these sections, the level of measurement appropriate for the latent constructs is nominal. On the other hand, the Rasch model was very capable of fitting those sections that had no misfitting items. These latent constructs have interval levels of measurement. Obviously, a priori considerations should dictate the level of measurement of any latent construct(s) being modeled. But in this case, as in many others, this knowledge may not be available. This example provides support and methods for LCA's use in helping to derive this information in these situations.

Future Directions

Only the Latent Class Model with the best goodness of fit statistics was chosen for analysis. An in-depth analysis of the various other reasonably fitting models is an area for possible future research.

The individual items could be tweaked to improve the fit of the models without eliminating the items. For example, some of the rating scales could be converged down to the best fitting number of categories. Or, some of the attitudes questions could be

inverted so that they have the opposite sentiment. This might help to eliminate the distortion in the second half of this section. Also, further investigation into the type of questions used in this section might show that there are primarily two types of attitudes questions. Thus these two different constructs might be scaleable in two separate sections.

This study was done using only four sections of data on one type of questionnaire. More research could be done using other data on different topics which use different formats. This would be useful in further justifying some of the new methods outlined within.

Limitations

None of the hybrid models were chosen for further investigation. There is little information to support the accuracy of these models. There is also little known about the implications of their use.

In contingency tables of enormous sizes such as the ones used in this study, cells often have observed frequencies of 0. The ln of 0 is an impossibility and causes problems in analyzing such data. There were many cells with 0 frequencies in the evaluation data. It is unknown how WINMIRA deals with these cells. This is taken to be a general weakness in this study as well as any other log-linear study with zero cells.

Although this is not a limitation, the specifications for the Estimation Maximization (EM) algorithm was set so that the accuracy criterion was =0.01, and the maximum number of iterations was set at 250. Some of the models did not converge at

250 iterations meaning that the accuracy of the fit statistics is greater than 0.01. Most of these models were hybrid models or LCAs of a high number of classes. Essentially, these models were the complex models having many parameters. Due to this complexity, lack of convergence, and the uncertainty of hybrid model results, all of these models were not investigated further.

APPENDIX: THE SURVEY QUESTIONNAIRE USED

BIBLIOGRAPHY

Anderson, T. W. (1954). On estimation of parameters in latent structure analysis. Psychometrika, 19, 1-10.

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

Barnard, B., & Haefele, D. (1993). Teachers' and principals' perceptions of teacher evaluation practices. Paper presented at the Annual Meeting of the Mid-Western Educational Research Association, Chicago.

Birnbaum, A. (1958). On the estimation of mental ability. (Series Rep. No. 15. Project No. 7755-23). Randolph Air Force Base, Texas: USAF School of Aviation Medicine.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley.

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, R. D., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM-algorithm. Psychometrika, 40, 443-459.

Bockenholt, U., & Bockenholt, I. (1990). Modeling individual differences in unfolding preference data: A restricted latent class approach. Applied Psychological Measurement, 14, 257-269.

Clogg, C. C. (1988). Latent Class Models for Measuring. In R. Langeheine & R. Rost (Eds.), Latent Trait and Latent Class Models. New York: Plenum Press.

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

Cronbach, L. (1984). Essentials of psychological testing (4^{th} ed.). New York: Harper & Row.

Dawis, R. V. (1987). Scale Construction. Journal of Counseling Psychology, 34, 481-489.

De Ayala, R. J. (1993). An introduction to polytomous item response theory models. Measurement and Evaluation in Counseling and Development, 25, 172-189.

Demming, W. E., & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. Annals of Mathematical Statistics, 11, 427-444.

Everitt, B. S., & Dunn, G. (1988). Log-Linear Modeling, Latent Class Analysis, or Correspondence Analysis which method should be used for the analysis of categorical data? In R. Langeheine & R. Rost (Eds.), Latent Trait and Latent Class Models. New York: Plenum Press.

Formann, A. K. (1984). Die latent-class-analyse. Weinheim: Beltz.

Goodman, L. A. (1974a). The analysis of systems of qualitative variables when some of the variables are unobservable. Part I-A modified latent structure approach. American Journal of Sociology, 79, 1179-1259.

Goodman, L. A. (1974b). Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika, 61, 215-231.

Green, B. F. (1951). A general solution of the latent class model of latent structure analysis and latent profile analysis. Psychometrika, 16, 151-166.

Guttman, L. (1944). A basis for scaling qualitative data. American Sociological Review, 9, 139-150.

Haberman, S. J. (1979). Analysis of qualitative data: Volume 2. New developments. New York: Academic Press.

Haertel, E. (1990). Continuous and discrete latent structure models for item response data. Psychometrika, 55, 477-494.

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

Kelderman, H. (1984). Log linear Rasch model tests. Psychometrika, 49, 223-245.

Kerlinger, F. N. (1986). Foundations of Behavioral Research (3rd ed.). Orlando: Harcourt Brace Jovanovich College Publishers.

Langeheine, R., & Rost, J. (1988). Latent Trait and Latent Class Models. New York: Plenum Press.

Langeheine, R. (1988). Developments in Latent Class Theory. In R. Langeheine & R. Rost (Eds.), Latent Trait and Latent Class Models. New York: Plenum Press.

Lawley, D. N. (1943). On problems connected with item selection and test construction. Proceedings of the Royal Society of Edinburgh, 62-A, 74-82.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Claussen (Eds.), Measurement and prediction: Studies in social psychology in World War II (Vol. IV). Princeton: Princeton University Press.

Lazarsfeld, P. F., & Dudman, J. (1951). The general solution of the latent class case. In P. F. Lazarsfeld (Ed.), The use of mathematical models in the measurement of attitudes. Santa Monica: RAND Corporation.

Lazarsfeld, P. F., & Henry, N. W. (1968). Latent structure analysis. Boston: Houghton Mifflin.

Linacre, J. M. (1994). Many-Facet Rasch Measurement. Chicago: MESA Press.

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

Lord, F. M. (1952). A theory of test scores. Psychometric Monograph, 7.

Lord, F. M. (1953a). An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. Psychometrika, 18, 57-75.

Lord, F. M. (1953b). The relation of test score to the trait underlying the test. Educational and Psychological Measurement, 13, 517-548.

Margenau, H. (1950). The nature of physical reality. New York: McGraw-Hill.

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149-174.

Masters, G. N. (1985). A comparison of latent trait and latent class analyses of likert-type data. Psychometrika, 50, 69-82.

McDonald, R. P. (1982). Linear versus non-linear models in item response theory. Applied Psychological Measurement, 6, 379-396.

McHugh, R. B. (1956). Efficient estimation and local identification in latent class analysis. Psychometrika, 21, 331-347.

McHugh, R. B. (1958). Note on "Efficient estimation and local identification in latent class analysis." Psychometrika, 23, 273-274.

Mislevy, R. J., & Bock, R. D. (1982). BILOG: Maximum likelihood item analysis and test scoring with logistic models for binary items. Chicago: International Educational Services.

Osgood, C. E., Suci, C. J., & Tannenbaum, P. H. (1957). The measurement of meaning. Urbana: University of Illinois Press.

Owen, R. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 70, 351-356.

Rasch. G. (1980). Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement, 17.

Stephenson, W. (1953). The study of behavior. Chicago: University of Chicago Press.

Thissen, D. M. (1982a). Marginal maximum likelihood estimation for the one-parameter logistic model. Psychometrika, 47, 175-186.

Thissen, D. M. (1982b). MULTILOG: Item analysis and scoring with multiple category response models. Chicago: International Educational Services.

Thurstone, L. L., & Chave, E. (1929). The measurement of attitude. Chicago: University of Chicago Press.

Torgerson, W. S. (1958). Theory and method of scaling. New York: Wiley.

Tucker, L. R. (1946). Maximum validity of a test with equivalent items. Psychometrika, 11, 1-13.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). LOGIST user's guide. Princeton, NJ: Educational Testing Service.

Wright, B. W., & Linacre, J. M. (1991). A user's guide to Big Steps. Chicago: MESA Press.

Wright, B. D. (1991). Rasch vs. Birnbaum. Rasch Measurement Transactions, 5, 178-179.

Wright, B. D. (1992). IRT in the 1990s: which model works best? Rasch Measurement Transactions, 6, 196-200.

Wright, B. D., & Stone, M. H. (1977). Best Test Design. Chicago: MESA Press.

Rost, J., & von Davier, M. (1994). A conditional item fit index for Rasch models. Applied Psychological Measurement, 18, 171-182.

von Davier, M. (1995). WINMIRA V1.68 User Manual. Kiel, Germany: Institute for Science Education.

Yamamoto, K. (1989). A Hybrid model of IRT and latent class models. (ETS No. RR-89-41), Princeton, NJ: Educational Testing Service.