PPROXIMITY MEASURES AND CLUSTERR ANALYSES IN MULTIDIMENSIONAL ITEM RESPONSE THEORY

By

JONG-PIL KIM

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling, Educational Psychology,
and Special Education

2001

ABSTRACT

PROXIMITY MEASURES AND CLUSTER ANALYSES
IN MULTIDIMENISONAL ITEM RESPONSE THEORY

By

JONG-PIL KIM

Developments in multidimensional item response theory (MIRT) have increased the development of dimensionality tools to assess the dimensional structure of item pools. This study investigated the comparative effectiveness of the hierarchical cluster analyses (HCA) when using non-parametric and parametric proximity measures for identifying the dimensional structure of data. The nonparametric approach is based on the assumption that the patterns of local independence in the conditional covariances can yield information about the dimensional structure, while the parametric approach is based on the angular distance (direction cosines) converted from direction estimated multidimensional discrimination parameter estimates.

Simulation studies were designed to determine if item difficulty, guessing, and ability levels influence the correct classifications of two approaches under various conditions. Different proximity measures, HCAs, sample

sizes, number of clusters, test length, and dimensional structures (approximate independent [APSS] and mixed structure [MS]) were considered as simulation factors.

Major findings of the study were: (1) CA using non-parametric similarity measures (especially the average method) were successful (higher than 90% recovery) for the APSS, but not successful for the MS. (2) CA using parametric similarity measures (especially Ward's method) were successful for both the APSS and the MS. (3) While difficulty and guessing levels did not affect the clustering results, ability levels (especially at lower ability) influenced the clustering results.

Since Ward's method with the angular distance yielded stable classifications under various test conditions, the parametric approach is recommended for analyzing the structure of test content. Wise use of proposed parametric and nonparametric approaches together can provide useful information about the dimensional structure. Because both methods work well with large sample sizes, they are useful for large-scale assessments such as standardized achievement/attitude tests. Caution should be used when a test is measuring relatively large number of diverse traits with small numbers of items related to each of them.

ACKNOWLEDGEMENTS

Today's accomplishment is primarily due to consistent support and unfailing contributions of my respected dissertation advisor Dr. Mark Reckase. I am greatly indebted to his guidance and inspiration that always have given me energy to finish all the details. His open-mindedness and rich experience in psychometrics and real life have been very precious to me over the years.

I am especially appreciative of Dr. Betsy J. Becker for her thoughtful contribution not only on my dissertation but also on my doctoral program itself. She is my academic guidance advisor, and without her sincere consideration I could not possibly finish all program requirements and many other things.

I also want to appreciate Dr. William Mehrens' sharing whole-life-experience on our academic area with me that might have influenced my academic pursuits and maybe my lifetime goals too. I feel very sorry for the other colleague doctoral candidates because he is no longer with us to teach and share with his deepest insights and widest academic scope. He will be always be remembered by those who had the privilege of working closely with him as a great scholar, intellectual and a friend to all.

I cannot ignore the sincere help from Dr. Ann Marie Ryan in Psychology department for her dedication in thoughtful reading and valuable comments on my dissertation. I do believe that her sincere concern about this dissertation has given me many fortunes.

In addition, I wish to extend my appreciation to Dr. Lois Roussos, Dr. S.W. Choi, and Dr. Yuan H. Li for generously providing their computer program codes to generate multidimensional item responses and for giving some valuable comments on the study. Also, I would like to acknowledge the financial supports for this dissertation provided via the Experimental Psychological Dissertation Awards and Spencer/Michigan State University Small Grant Research Award. I was so honored.

In closing, a wholehearted and yet indescribable feeling of warmth, gratitude, and love is expressed in simple words "thank you" to my parents who supported me throughout every moment of the entire my life until today. To my lovely wife, Jae-Moon Kim, who has patiently supported me in countless ways with love and caring ever since our marriage, I am totally indebted. I do believe that all my appreciative expression to all people could be replaced with these three words, "Thank you, God."

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

Early item response theory (IRT) models were based on the assumption of unidimensionality. When more than one trait accounts for test performance and the test items are sensitive to more than one trait, the matrix of item responses is multidimensional and a multidimensional IRT (MIRT) model is appropriate for describing the characteristics of the data. Several authors (Drasgow & Lissak, 1983; Harrison, 1986; McKinley & Kingston, 1988) have argued that achievement tests are likely to yield multidimensional item response data because of the several content areas that are included in the particular subject. This also applies to psychological measures like personality or attitude scales.

Among recent increasing concerns about multi-dimensional aspects of test data, one important issue is the number of cognitive dimensions needed to respond correctly to the items in an item pool. According to Lord and Novick (1968), dimensionality is defined as the total number of abilities required to satisfy the assumption of local independence in an IRT sense. The latent trait space is unidimensional if and only if a single trait is

sufficient to account for the persons' performance on the test; otherwise the latent trait space is multidimensional (Zeng, 1989).

Determining the number of dimensions that is needed to summarize a set of data has been a problem since the time of Spearman (1904) who argued that performance on sets of tests could be explained by individuals' levels on general and specific traits. Since that time, many different procedures have been suggested for determining the number of dimensions needed to describe a set of data. These include the number of eigenvalues greater than 1.0 (Kaiser, 1970), a discontinuity in the curve formed by a scree plot (Cattell, 1978), parallel analysis (Horn, 1965), tests of fit based on maximum likelihood estimation procedures (Bock & Aitkin, 1981), and the DETECT procedure (Kim, 1994) based on the work of Junker and Stout (1994).

Recently, Roussos, Stout, and Marden (1998) presented three, dimensionally-sensitive, proximity measures (Pccor, Pccov, and P_{MH}) based on Roussos's (1995) and showed their use with hierarchical cluster analysis (HCA) in identifying dimensions (i.e., Stout, Habing, Douglas, Kim, Roussos, & Zhang, 1996). These three proximity measures are based on the weak principle of local independence (McDonald, 1981). The general idea is that the patterns of misfit in the

residual covariances after fitting a uni-dimensional model can yield information about the dimen-sionality of the item response data. Roussos et al. (1998) believe that these proximity measures are sensitive to the number of dimensions in an item response data matrix. Since these proximity measures do not use an item response model, they are often called nonparametric approaches.

As alternative approach was suggested by Miller and Hirsh (1992) who argued for a proximity formulation based on the multidimensional discrimination indices from the compensatory normal-ogive multidimensional item response model as estimated by NOHARM (Fraser, 1988). The procedure they used was to convert discrimination parameter estimates (MDISC: Reckase, & McKinley, 1991) to direction cosines and use those cosines to specify directions of measurement in a multidimensional space. The angles between item directions were cluster analyzed to determine unique dimensions. They then group together items with similar orientations in the θ space. For instance, two items with the same relative pattern of discriminations on the coordinate axes of the space have identical angles with the axes and measure the same composite of the abilities represented in the space. The direction cosines are not influenced by item difficulty, and discrimination is represented only in a

relative sense. Any given pair of items can differ considerably in difficulty and absolute discrimination and still have identical direction cosines (Miller & Hirsh, 1992, p. 210). Since this method uses an item response model, it is often called a parametric approach.

1.2 Purpose of the Study

The development and investigation of dimensionality assessment tools that have practical implications in measurement applications has become an important issue. The determination of the dimensional structure underlying a set of test items as well as the determination of the number of dimensions in a test are important to test developers and users. An incorrect specification of the dimensionality underlying a set of item responses can lead to serious problems in parameter estimation, test construction, test equating, and interpretation and use of test scores (Choi, 1997, p.6).

As Miller and Hirsh (1992) state, more research is needed to explore more fully the relationship between true dimensionality, the number of dimensions extracted, and the stability of the clustering procedure. Roussos (1995) compared two different proximity measures computed by non-parametric and parametric approaches in a real setting, and

concluded that two proximity measures have similar clustering results. However, he did not explain how similar they are, or how different they are. To answer those questions, a simulation study might have been implemented. The determination of the proximity matrix is usually a nontrivial task and it can have great impact on the results of the cluster analysis. A way of determining the proximities between the objects must be found such that this proximity is in harmony with the substantive objective of the cluster analysis. Furthermore, it is important to understand the factors that affect the clustering results, since cluster analysis is a sensitive classification tool.

The main purpose of this study is to compare the effectiveness of hierarchical cluster analyses with two different proximity measures (parametric and nonparametric) to identify the dimensional structure needed to model test data. A second goal is to determine the effect of guessing and different examinee ability levels on the clustering results in identifying the dimensional structure.

CHAPTER 2

LITERATURE REVIEW

2.1. Multidimensional Item Response Theory and Parametric Proximity Measures

Depending on a psychometrician's background, multidimensional item response theory (MIRT) can be considered either a special case of factor analysis or structural equating modeling, or an extension of unidimensional item response theory (IRT) (Reckase, 1997a, p. 25). In the late 1970s and early 1980s, a number of researchers were actively working to develop practical MIRT models.

Sympson (1978) proposed the following model with the probability of a correct response to item j by person i defined by a multidimensional three-parameter model.

$$P(x_{ij}=1|\theta_j,a_i,b_i,c_i)=c_i+(1-c_i)\prod_{k=1}^{m}\frac{\exp[a_{ik}(\theta_{jk}-b_{ik})]}{1+\exp[a_{ik}(\theta_{jk}-b_{ik})]}, \quad (1)$$

where x_{ij} is the score (0 or 1) on item i by person j, $P(x_{ij} = 1|\theta_j, a_i, b_i, c_i)$ is the probability in an m-dimensional space of a correct response to item i by person j, θ_j is an m-dimensional vector of latent abilities, a_i is an m-dimensional vector of discrimination parameters, b_i is a

vector of difficulty parameters, and c_i is a single pseudo-guessing parameter. The subscript k indicates kth dimension. Thus a_{ik}, θ_{ik}, and b_{ik} represent a vector of discrimination, trait, and difficulty values on the kth dimension respectively. This model is sometimes called the partially compensatory or non-compensatory model because an increase in θ on one dimension has only a minor effect on overcoming a deficit on another dimension (Reckase, 1997a, p.32).

The multidimensional two-parameter logistic IRT model proposed by McKinley and Reckase (1983), and Reckase and McKinley (1991) is defined as

$$P(x_{ij}=1|\theta_j,a_i,d_i)=\frac{\exp\left(\sum_{k=1}^{m}a_{ik}\theta_{jk}+d_i\right)}{1+\exp\left(\sum_{k=1}^{m}a_{ik}\theta_{jk}+d_i\right)}, \qquad (2)$$

where d_i is a parameter related to the difficulty of item i and the other symbols are defined above. This model is compensatory because the person parameters in the exponent of Equation (2) are additive over dimensions. It allows high levels on one dimension to compensate for low levels on other dimensions in arriving at the probability of correct response to an item.

The multidimensional compensatory three-parameter logistic IRT model based on Reckase (1985) is defined as

$$P(x_{ij}=1|\theta_j,a_i,d_i,c_i)=c_i+(1-c_i)\frac{\exp\left(\sum_{k=1}^{m}a_{ik}\theta_{jk}+d_i\right)}{1+\exp\left(\sum_{k=1}^{m}a_{ik}\theta_{jk}+d_i\right)}. \quad (3)$$

Item discrimination is related to how quickly the probability of a correct response changes as a function of θ. In unidimensional IRT, the item discrimination is proportional to the slope of the item characteristic curve at the point of inflection where the slope is steepest. In MIRT models, item discrimination parameters are given by the elements of the a-vector. Reckase and McKinley (1991) defined a multidimensional discrimination parameter (MDISC) for item i as

$$MDISC_i = \sqrt{\sum_{k=1}^{m}a_{ik}^2}, \quad (4)$$

where m = the number of dimensions in the ability space, and a_{ik} = the discrimination of item i on dimension k. *MDISC* is proportional to the slope of the item response function at the point of steepest slope and it is therefore analogous to the unidimensional discrimination parameter

(Carlson, 1987; Reckase & McKinley, 1991). Another important element in the MIRT model is the location of an item in the latent ability space. This location can be represented by a vector with the degree of $MDISC_i$ as the length of the vector for item i. The direction of the vector is calculated by the direction cosines, which use the following expression

$$\cos\alpha_{ik} = \frac{a_{ik}}{\left(\sum_{k=1}^{m} a_{ik}^2\right)^{1/2}}, \tag{5}$$

where the a_{ik} are elements of the vector a_i given in Equation (2) or (3).

Reckase (1985) defined a multidimensional item difficulty parameter (MDL) as both the direction in the space given by Equation (5) and the distance from the origin to the point of steepest slope given by

$$\mathrm{MDL}_i = \frac{-d_i}{MDISC}. \tag{6}$$

Reckase, Ackerman, and Carlson (1988) showed that the direction of an item vector indicates the weighted composite of abilities best measured by the items and that

items with the same or similar direction cosines or angles in the latent space measure the same or similar weighted composites of abilities represented by that space. When there is more than one distinct set of items with similar direction cosines in a given latent space, each set can be treated as a different unidimensional composite of the abilities represented in the space. The amount of spread among the vectors that compose a set reflects the degree to which unidimensionality does not hold for that set. Thus, the evaluation of direction cosines or angles of item vectors in a MIRT analysis provides a means of identifying similarly functioning sets of items in terms of what they measure and in that sense can provide useful insights into the structure of a test (Miller & Hirsh, 1992, pp. 196-197).

NOHARM (Fraser, 1988) is a program for estimating the a_{ik} and d_i parameters (in Equation (2) and (3)) by fitting the multidimensional normal ogive model by a least-squares procedure. This program requires c_i parameters as input, so they need to be obtained by some other method such as PC-BILOG (Mislevy & Bock, 1982). Angular distances α_{ij} between item vectors defined by MDL and MDISC are obtained from

$$\cos\alpha_{ij} = \frac{a_i' \cdot a_j}{|a_i||a_j|}, \quad (7)$$

where a_i and a_j are the vector of discriminations for item i and j. NOHARM provides options for rotating the solution to different orientations including unrotated, varimax (orthogonal) rotated, and promax (oblique) rotated analyses. Each analysis provides discrimination estimates using a different parameterization.

Miller and Hirsch (1992) and Reckase (1997c) showed how to use the matrix of angular distances as a dissimilarity matrix for hierarchical cluster analysis. They showed an example using the angular distance with the complete linkage method to analyze the test structure of the 40 multiple-choice items of the P-ACT+ Mathematics Test. This test was measuring achievement in the areas of pre-algebra, elementary algebra, coordinate geometry, and plane geometry at the high-school level. They used 2-dimenisonal and 6-dimensional solutions in NOHARM and concluded that a 6-dimensional solution gave a better representation of the test structure. However, the reason they used a 6-dimensional solution was because it was the highest-dimensional solution supported by NOHARM at that time. At this time, NOHARM supports up to a k-1 dimensional

solution for k, the number of items. If higher number of dimensional solution were used, a better solution might have been obtained.

Reckase (1997b) tried to compare constructs underlying portfolios (PASSPORT) and the ACT Assessment using the angular distance with HCA. The cluster analysis with a parametric approach effectively showed the difference in the types of skills and knowledge that are assessed. Lecht and Miller (1992) also used this parametric approach to evaluate a two-stage process for modeling composite latent traits.

The angular distance between two items is illustrated in Figure 2.1. The arrows indicate the projection of the direction of steepest slope of the item response surfaces

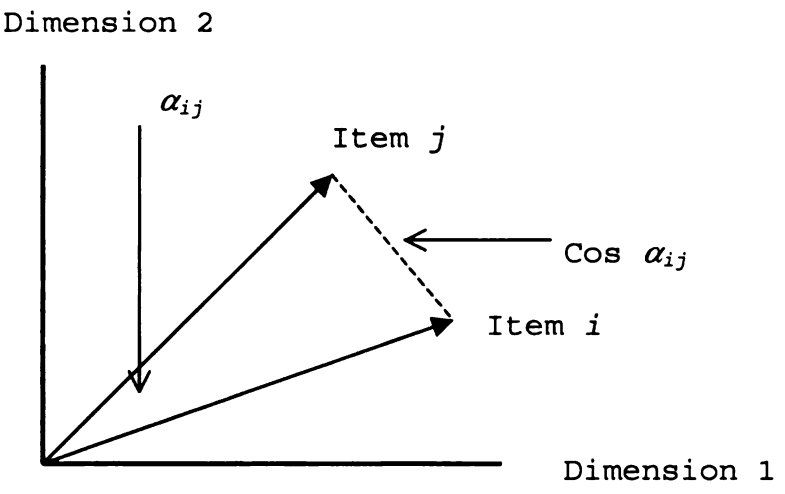


Figure 2.1. The Angular Distance in Two-Dimensional Space

for item i and j, onto the θ_1 and θ_2 plane. The angular distance provides a means of identifying similarity and dissimilarity of sets of items in terms of what they measure, and can provide useful information about the structure of a test (Miller & Hirsh, 1992). If Cos α_{ij} is close to 1.0, then item i and j measure the same traits (or belong to the same dimensions. If Cos α_{ij} is close to 0.0, then item i and j measure different traits (belong to different clusters).

At least 2000 examinees are needed to obtain satisfactory MIRT parameter estimates. This sample size limits the use of the parametric approach to large national testing programs (e.g., ACT, ASVAB), statewide testing programs, large scale selection assessment tools, etc.

2.2 Nonparametric Proximity Measures

As opposed to the above parametric approach, Roussos et al. (1998) proposed three item proximity measures, which are often called nonparametric proximity measures. The general idea behind these proximity measures is that the patterns of misfit in the residual covariances after fitting a unidimensional IRT (UIRT) model, or equivalently, the patterns of local independence in the conditional

covariances, can yield information about the dimensionality structure in the items (Chen & Thissen, 1994; Hambleton & Rovinelli, 1986; McDonald, 1982; Roznowski, Tucker, & Humphreys, 1991; Yen, 1984). The proximity measures are based on the contingency table in Table 2.1 for scores of examinees on two items after the examinees have been partitioned into groups of approximately equal ability.

For a test containing N items, the derivation of a proximity measure for any item pair, i and j, could begin with a set of $N-1$ contingency tables, one 2X2 table for each of the possible number-correct scores, s, that an examinee could have on the remaining N-2 items. X_i is the score on item i and X_j is the score on item j. A_s is the number of examinees who answered both correctly, B_s is the number of examinees who answered i correctly but j incorrectly, C_s is the number of examinees who answered j correctly but i incorrectly, and D_s is the number of examinees who answered both items incorrectly. Also $n_s = A_s + B_s + C_s + D_s$. Three proximity measures that may be useful

Table 2.1. Contingency Table for Two Dichotomous Items

		X_j	
		1	0
X_i	1	A_s	B_s
	0	C_s	D_s

for identifying multidimensionality are written as follows:

$$\text{Pccor} = \sqrt{2(1-\frac{1}{\sum n_s}\sum_{s=0}^{N-2} n_s COR_s)}, \tag{8}$$

$$\text{Pccov} = -\frac{1}{\sum n_s}\sum_{s=2}^{N-2} n_s COV_s + \text{constant}, \quad \text{and} \tag{9}$$

$$\text{P}_{\text{MH}} = -\log\left(\frac{\sum_{s=0}^{N-2}\frac{A_s D_s}{n_s}}{\sum_{s=0}^{N-2}\frac{B_s C_s}{n_s}}\right) + \text{constant}, \tag{10}$$

where a constant is added to two of the proximity measures to ensure that they are greater than or equal to zero, COV_s is the observed covariance between two items computed from those examinees with score s, and COR_s is the product-moment correlation corresponding to COV_s. P_{MH} is based on the Mantel-Haenszel log-odds ratio (Holland & Rosenbaum, 1986; Mantel & Haenszel, 1959; Rosenbaum, 1984).

For a test composed of items sensitive to more than one ability, if two items measure the same dimension, positive local dependence occurs when conditioning on the latent multidimensional composite ability. When two items measure different dimensions, negative dependence occurs. As Roussos et al. (1998) stated

"If two items measure the same dimension, say, math, then the examinees who got the first math item right will tend, as a group, to have a higher probability of getting the second math item right than will the entire group of examinees. …. Thus, the covariance conditional on the composite ability is seen to be positive when the two items measure the same dimension." (p. 7)

HCA/CCPROX (Roussos, 1992) consists of two main computer programs CCPROX and HCA. CCPROX is designed to produce nonparametric proximity measures (Pccor, Pccov, and P_{MH}) for all item pairs on a dichotomously scored test. CCPROX does not calculate proximity between two items if there are any empty cells in the contingency table (See Table 2.1). In such cases, the easier of the two items is ignored and the proximity calculation continues. Thus monitoring the ignored item(s) is important for understanding the HCA outcome. The HCA program is designed to analyze data using hierarchical cluster analysis. The HCA methods implemented in the HCA program are the single, complete, UPGMA (unweighted pair-group method of averages), WPGMA (weighted pair-group method of averages), mean dissimilarity, sum of squares, Ward, UPGMC (unweighted pair-group method of centroids), WPGMC (weighted pair-group method of centroids), and flexible method.

Stout et al. (1996) showed how to apply hierarchical cluster analysis (HCA) to the proximity measures to detect

the dimensionality in a test. They used HCA/CCPROX with DIMTEST (Nandakumar & Stout, 1993; Stout, 1987) or DETECT (Kim, 1994; Zhang & Stout, 1995, 1996). They use HCA/CCPROX with DIMTEST or DETECT because HCA/CCPROX does not offer statistical tests. After analyzing three administrations of the LSAT (Law School Admission Test: December, 1991; June, 1992; and October, 1992), Stout et al. (1996) concluded that the combined usage of HCA/CCPROX with DIMTEST and DETECT allowed for testing unidimensionality and for estimating the amount of multidimensionality presented in a set of test data.

Reckase (1994) emphasized that the appropriate dimensionality for an analysis depends on the goal of the analysis. However, there is still a question about 'which one is more effective in clustering items into given dimensions (clusters) under simulated conditions?' Even though Roussos (1995) compared them and concluded that these two approaches showed similar clustering results, he did not compare them using simulated data, but compared them using a real data set. The first research purpose for this study is to compare two different types of proximity measures (parametric vs. nonparametric) for use with a clustering procedure for analyzing simulated multi-dimensional data.

2.3 Dimensionality and Affecting Factors

There has been a question about the relationship between dimensionality and difficulty. As Ackerman (1989) states, difficulty seems to be confounded with dimensionality. For instance, order analysis methods based on Guttman scaling tend to confound difficulty level with dimensionality. That is, items of similar difficulty level but measuring different traits tend to be assigned to the same dimension (Roznowski, Humphreys, & Davy, 1994; Wise, 1983; Wise & Tatsuoka, 1986).

Oltman, Stricker, and Barrows (1988; 1990), using multidimensional scaling techniques, found that the easier items in each section of a test tended to define clusters and the more difficult items did not fit well into any of the dimensions identified in the test. Their results indicated the dimensionality of the Test of English as Foreign Language (TOEFL) depends on the level of English proficiency of the examinees, with more salient dimensions found in the least proficient populations of test takers. They concluded that the easy and difficult items were different in their ability to measure overall language proficiency and specific language skills, with easy items better suited for diagnostic purposes such as measuring specific language skills, and difficult items better

measures of general proficiency and, therefore, more useful for global screening purposes (Olson, Scheuneman, & Grima, 1989). Because examinees make errors on difficult items, these items might be expected to be more likely to cluster in ways that depend on errors.

Roznowski, Tucker, and Humphreys (1991) pointed out that the sampling errors of individual tetrachoric correlations vary greatly with item difficulty levels and concluded from their simulations that linear factor analysis on the sample tetrachoric correlation matrix is not dependable for any set of items that has a wide range of difficulties. In such a case, the use of the usual factor analysis criteria for determining the number of factors tends to point to more dimensions than are actually present (i.e., Hambleton & Rovinelli, 1986; Nandakumar, 1994).

As many researchers (e.g., Ackerman, 1992; Kok, 1988; Mazor, 1993) have indicated, judgments about the dimensions assessed by items are meaningful only with respect to a specific population. A test may be sensitive to *n* dimensions in one population, and be *n+1* dimensions in another. Thus dimen-sionality of a given data set is really a function of both the examinee sample and the item characteristics. Even if item difficulty, the ability level

and variability of the examinee population, and guessing are irrelevant to the assessment of content, these variables may influence the results of item response analyses (Sireci & Geisinger, 1992). It may not be easy to distinguish the real dimensionality from the influence of the above variables.

When Roussos et al. (1998) compared his new proximity measures with three classical proximity measures (P_{A+D}, P_A, and P_{cor}, see Cronbach & Gleser, 1953; Gower & Legendre, 1986), they assumed that the classical proximity measures would fail to distinguish dimensionality and difficulty while their nonparametric proximity measures would be able to identify the exact dimensions without any interference by difficulty. On the other hand, Miller and Hirsch (1992) also stated that the direction cosines are not influenced by difficulty. They implied the angular distance proximity is free from item difficulty effects. However, there is no study investigating if the parametric approach is unaffected by difficulty, guessing and different ability levels. The second research purpose of this study is to determine if ability levels and guessing affect the results of clustering models when using HCA methods with parametric and nonparametric proximity measures.

2.4 Cluster Analyses

The wide variety of cluster analysis (CA) methods has led some authors to classify them into several families of methods (i.e., Blashfield & Alenderfer, 1978; Lorr, 1983). One of the most-used methods, the hierarchical cluster analysis (HCA), begins with a matrix of similarities or dissimilarities between each pair of cases. Based on the proximity matrix and a rule for determining similarity between clusters, hierarchical agglomerative methods consist of ($n-1$) stages, where at each stage the two most similar clusters are joined together to form a new cluster. A hierarchy may be considered as a family of nested multilevel classes. Thus, a hierarchical analysis combines individual cases (in this study, items) into small clusters, then combines small clusters into larger ones, and at ever higher levels keeps combining clusters until all have been subsumed under one grand heading. 'Agglomerative' means that it begins with N entities that are sequentially merged at successive levels until all are included (Lorr, 1983).

Five commonly used HCA methods are Single Link (SLINK: Florek, Lukaszewicz, Perkal, Steinhaus, & Zubrzycki, 1951), Complete Link (CLINK: McQuitty, 1960), unweighted pair-group method of averages (UPGMA: Sokal & Michener, 1958),

weighted pair-group method of averages (WPGMA: McQuitty, 1966), and Ward's minimum variance clustering method (Ward's method: Ward, 1963). Different linkage rules often lead to different clusterings of cases. Roussos (1995) indicated that UPGMA performed the best of the HCA methods with WPGMA second best, while Miller and Hirsch (1992) used CLINK methods.

Single Link starts with finding and merging the two clusters i and j (treating the N entities as N clusters) within the set that are separated by the smallest distance d_{ij}. The distance between the new cluster k and some other cluster h is defined as the minimum distance between an observation in one cluster and an observation in the other cluster. Thus this method tries to find $d_{hk} = min(d_{ik}, d_{jk})$, where d_{hk} is the distance between the closest members of clusters h and k.

Complete Link procedure defines the distance between clusters as the distance between their most remote pair of entities. Thus the distance d_{hk} is equal to max (d_{ik}, d_{jk}), where max means the larger of the distances compared and d_{hk} represents the separation between the most remote or distant members of clusters h and k. In Average Link methods (UPGMA and WPGMA), each member of a cluster has a smaller average dissimilarity with other members of the

same cluster than with members of any other cluster. The distance between clusters is defined as the average of the distance between all pairs of entities in the two clusters. The distance between two clusters in Ward's method is the ANOVA sum of squares between the two clusters, added up over all the variables (items). At each generation, the within-cluster sum of squares is minimized over all partitions obtainable by merging two clusters from the previous generation (Lorr, 1983; SAS, 1990, pp.530~536).

CHAPTER 3

METHODS

3.1 Study I

The first purpose of this study is to compare the effectiveness of hierarchical cluster analyses with two different proximity measures from parametric and non-parametric approaches to identify the dimensional structure needed to model test data. To seek an answer to this, a simulation study was conducted in a way similar to Roussos et al. (1998, pp. 10-14). The simulation study involved four major variables: proximity measures, HCA methods, number of examinees, and dimensional structure. The simulation model corresponding to each combination of variables was run 100 times. Each variable is described in detail below.

3.1.1 Simulation Factors

(1) Proximity Measures

The angular distance was used as a parametric proximity measure. One proximity measure, Pccor, was used for the nonparametric approach. Pccor is the most natural proximity measure because artificial constants need to be added to the other two proximity measures (Pccov & P_{MH}) to ensure positive proximities for all the item pairs. In

addition, Pccor is one of most effective proximity measure among three proximity measures (Roussos et al., 1998).

(2) HCA Methods

Four HCA methods were included in this study: single link, complete link, unweighted pair-group method of average (UPGMA), and the Ward method. WPGMA was not used because of its resemblance to UPGMA. Because Ward's method has been found to be more stable than other clustering methods (i.e., Jain & Dubes, 1988), it was included in this study.

(3) Number of Examinees: 1000 and 3000.

It has been known that MIRT estimates are stable with 2000 examinees. Thus more than that sample size (3000) and less than that sample size (1000) were considered.

(4) Dimensional Structure

Four different dimensional structures were simulated as follows.

(a) Number of Dimensions: 2- and 3-dimensions.

The latent dimensionality was represented by an ability vector θ, for instance, (θ_1, θ_2) in the two-dimensional case (2-D). These vectors were modeled as

multivariate normal with mean vector of 0 and standard deviations of 1.

(b) Number of Items: 20, 30, & 60 for 2-D; and 30 & 60 for 3-D.

(c) Number of Clusters

To decide upon the number of clusters in a test, several kinds of structure are explained. First, simple structure (SS) refers to a test that can be divided into clusters, each of which corresponds to a separate test dimension (all angles for each item are 0° or 90° with each dimension). SS was not considered in this study because it is hard to obtain dimensionally pure sets of items.

Second, approximate simple structure (APSS) was considered. Roussos et al. (1998) inspected both SS and APSS. APSS refers to a test with each item in a cluster allowed to have small amounts of discrimination on the dimensions of the test while the items in the cluster have their highest discrimination on a single dimension. Figure 3.1 shows an example of the APSS in three-dimensional space with 30 items. Each set of 10 items corresponding to a dimension k is allowed to randomly fall within 15° of the k^{th} axis. Those items corresponding to other dimension(s) have in $(90^\circ - \alpha_{ik}) \leq \alpha_{ik} \leq 90^\circ$, where α_{ik} indicates the angle between item i and dimension k. For a given α_{ik} (randomly

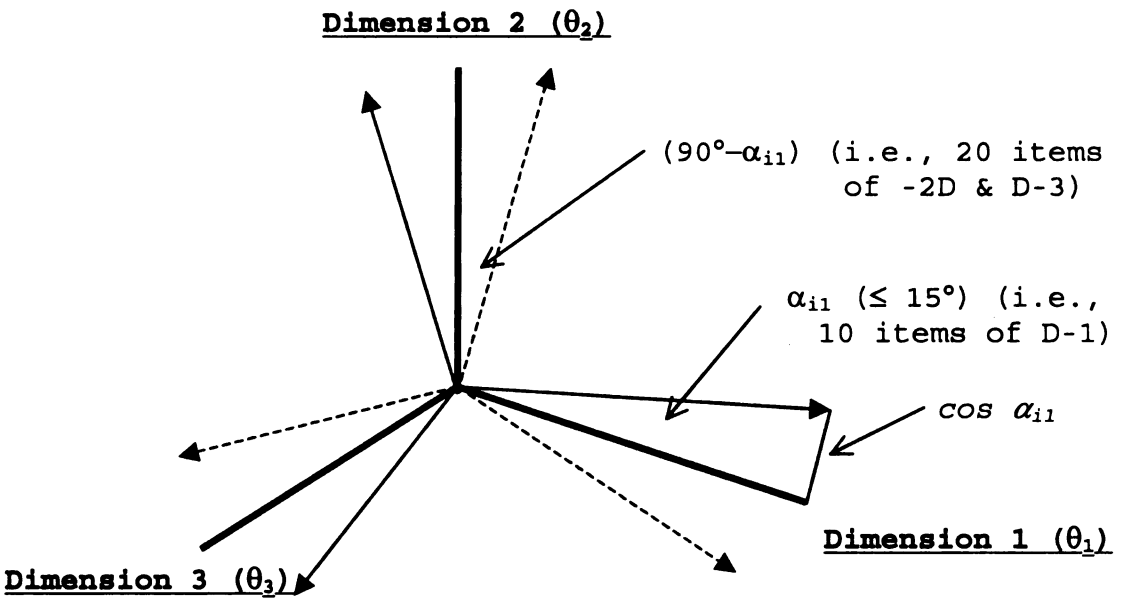
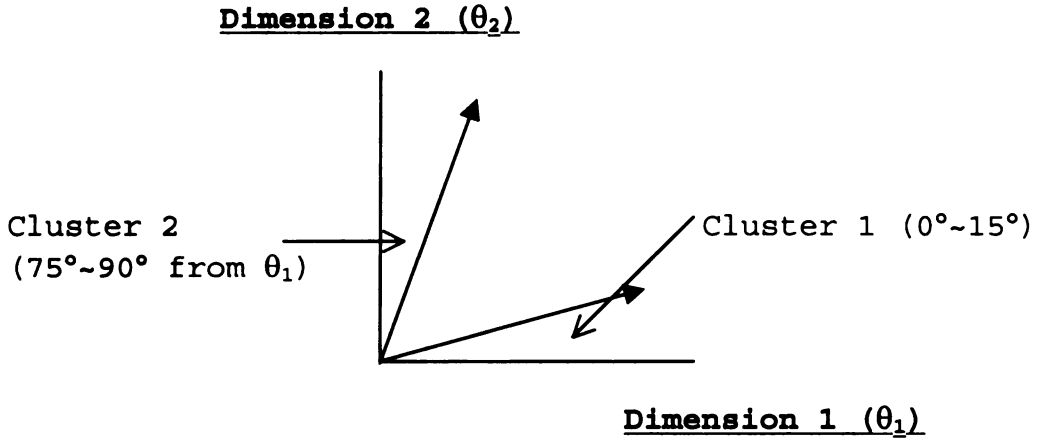


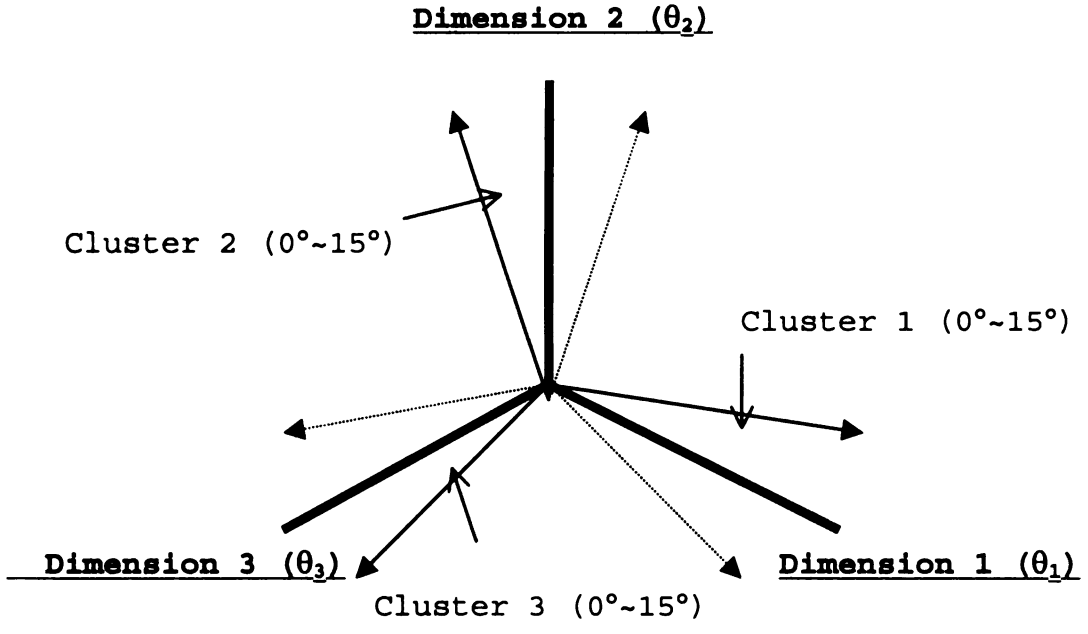
Figure 3.1. An Example of APSS in Three-Dimensional Space (30 Items)

drawn), a_{ik}, elements of the vector of a_i given in Equation (5) can be computed using $cos\ \alpha_{ik}$ multiplied by *MDISC*.

Figure 3.2 illustrates possible dimensional structures of 2-D and 3-D tests when using APSS. Each case has the same number of clusters as dimensions. Table 3.1 shows the number of items per cluster and angles from each dimension (trait) when using APSS. Even though the APSS method Roussos et al. (1998) used has some problems, the above conditions were included in this study so the results of this study could be compared with those of Roussos' study. They simulated only tests that had the same number



(a) 2-Dimensions



(b) 3-Dimensions

Figure 3.2. APSS for Two- and Three-Dimensional Tests

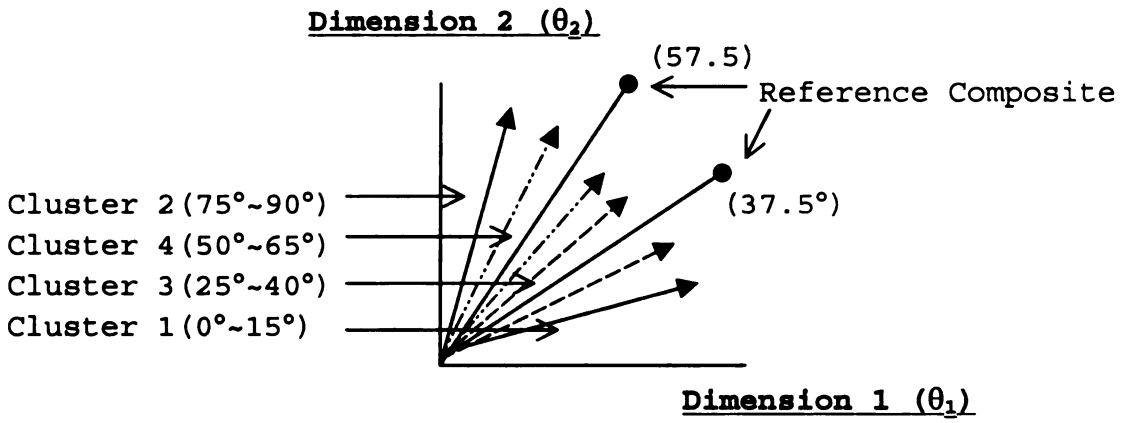
Table 3.1. Number of Items per Cluster and Angles from Dimension for APSS

	2 Dimensions			3 Dimensions			
Cluster	No. of Items	Angles from Trait 1	Trait 2	No. of Items	Angles from Trait 1	Trait 2	Trait 3
1	10,20,30[1)]	0~15	75~90	10,20[2)]	0~15	75~90	75~90
2	10,20,30	75~90	0~15	10,20	75~90	0~15	75~90
3				10,20	75~90	75~90	0~15

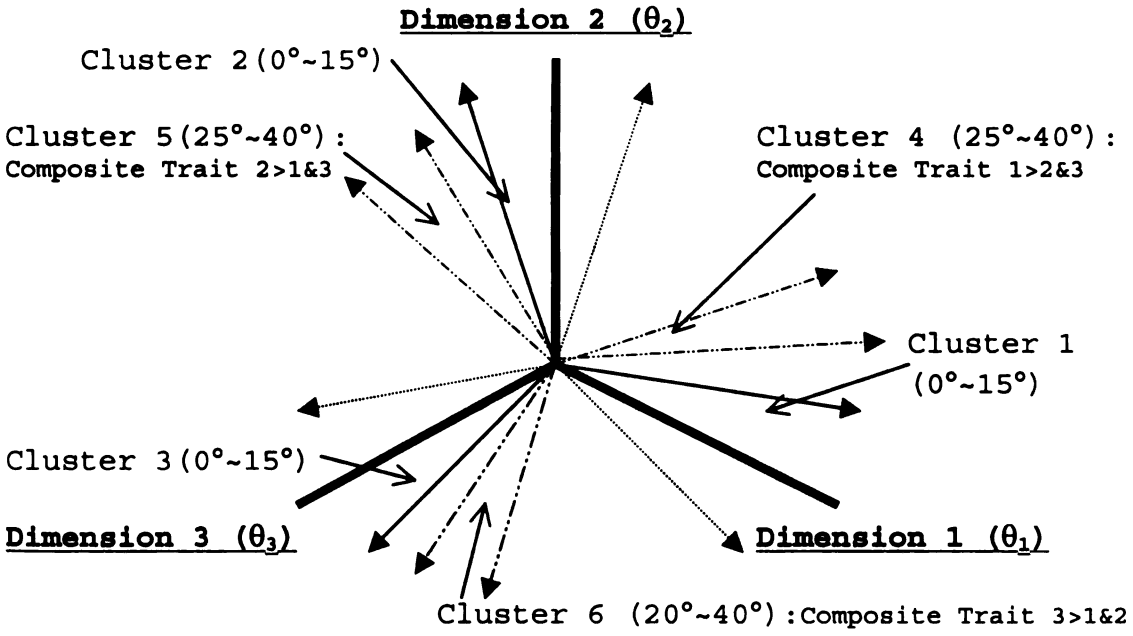
1) 10,20,30 under 2-D indicate the number of items in a cluster, so that test length would be 20, 40, & 60.
2) 10,20 under 3-D indicate the number of items in a cluster, so that test length would be 30 & 60.

of clusters as dimensions. For example, the condition for 2-D assumed that a test they were modeling was measuring only two different traits.

In reality, however, tests are not usually constructed to measure two or three independent traits. Rather, tests also measure some composite of traits as well. Thus, this study considered tests that had more clusters than the numbers of dimensions (traits). This kind of structure is often called mixed structure (MS: i.e., Kim, 1994). More specifically, conditions were added where a test modeling 2-D had 4 clusters (2 pure traits and 2 composite traits) and a test modeling 3-D had 6 clusters (3 pure traits and 3 composite traits). The MS conditions (2-D and 3-D) are illustrated in Figure 3.3. This clustering of items into subtests in terms of their directions of maximum information is related to Wang's (1986) concept of the reference composite. A reference composite is actually a



(b) 4 Clusters for 2-D Test



(b) 6 Clusters for 3-D Test

Figure 3.3. MS for Two- and Three-Dimensional Tests

weighted composite of the underlying multiple dimensions. The weights determining the direction (i.e., the relative proportion of two traits being measured) of the reference composite are a function of the MDISCs and the variance-covariance matrix of the underlying multi-dimensional ability distribution (Ackerman, 1992, p. 72). This reference composite is the unidimensional latent trait metric axis. In a cognitive or psychological sense, the composite traits being measured by each cluster may be multidimensional, depending on the orientation of the reference composite in the latent space. However, in a psychometric sense, a unidimensional IRT model should fit the data, because the clustered items all provide the maximum amount of discrimination in the same general direction (Luecht & Miller, 1992, p. 28). For a 2-D test with 4 clusters, the two underlying reference composites were oriented, respectively, at 32.5° and 57.5° from dimension 1 with a range of 15° around the dimension.

The numbers of clusters with specific angles from each dimension are presented in Table 3.2 for 2- and 3-D tests with the MS. For instance, in a 2-D test of 20 items with 4 clusters, 5 items (cluster 1 and 2) are measuring trait 1 and trait 2 respectively. The next five items (cluster 3) are measuring a composite trait of 1 and 2 with more

Table 3.2. Number of Items per Cluster and Angles from Dimension for MS

	2 Dimensions			3 Dimensions			
	No. of	Angles from		No. of	Angles from		
Cluster	items	Trait 1	Trait 2	Items	Trait 1	Trait 2	Trait 3
1	5,10,15[1)]	0~15	75~90	5,10 [2)]	0~15	75~90	75~90
2	5,10,15	75~90	0~15	5,10	75~90	0~15	75~90
3	5,10,15	25~40	50~60	5,10	75~90	75~90	0~15
	(Composite trait 1>2)						
4	5,10,15	50~60	25~40	5,10	25~40	50~60	50~60
	(Composite trait 1<2)				(Composite trait 1>2&3)		
5				5,10	50~60	25~40	50~60
					(Composite trait 2>1&3)		
6				5,10	50~60	50~60	25~40
					(Composite trait 3>1&2)		

1) 5,10,15 under 2-D indicate the number of items in a cluster, so that test length would be 20, 40, & 60.
2) 5,10 under 3-D indicate the number of items in a cluster, so that test length would be 30 & 60.

weighting on trait 1. The last five items (cluster 4) are measuring another composite trait of 1 and 2 with more weighting on trait 2. This test modeling emulates that of the 40 multiple-choice items of the P-ACT+ Mathematics Test for 10th grade (Miller & Hirsh, 1992; Reckase, 1997c).

For another example of a 3-D test of 30 items with 6 clusters, the first three five-item sets (clusters 1, 2, and 3) are measuring trait 1, 2 and 3. Next, three more sets of five items (clusters 4, 5, and 6) are measuring a composite trait of 1, 2, and 3, respectively, with different weighting on each trait. The items in cluster 4,

for example, measure θ_1 with relatively more weighting and θ_2 and θ_3 with lower weightings.

(d) Correlation between Dimensions

When using APSS (when the number of clusters is equal to the number of dimensions), a .7 inter-dimension correlation was used. Most simulation studies in MIRT have used correlation of .2 to .7 between dimensions (i.e., Batley & Boss, 1993; Hambleton & Rovinelli, 1986; Nandakumar, 1994; Roznowski, Tucker, & Humphreys, 1991; Yen, 1984;). As the correlation between dimensions is increased, clustering items into proper clusters is more difficult. Thus, the highest correlation, .7, was selected to investigate the effectiveness of parameter and non-parametric proximity measures under an extreme condition (highly correlated dimensions).

When 4 and 6 clusters were in the 2- and 3-D case respectively (when using the mixed structure), the correlations between dimensions (θs) were set at 0.0. This was because the specified angles between the item clusters reflected the correlation between dimensions. By doing so, the correlations among composite traits (for 2-D and 3-D tests) could have been kept between .65 and .75, which are realistic values for real tests. In addition, the

correlations between dimensions ("independent" traits) were between .52 and .64 for the 2-D test, and between .31 and .42 for the 3-D test.

3.1.2 Item Parameters and Data Generation

Roussos et al. (1998) used five types of items to check the ability of the proximity measures to identify similarity of dimensionality rather than similarity of difficulty (Table 3.3), as defined by $MDISC_i$ and MDL_i parameters. These parameters were repeated to get the designed test length for each dimension.

Given the structures in Table 3.1 and 3.2, the angles (α_{ik}) corresponding to dimension (θ_k) were randomly drawn from a uniform distribution (see Figure 3.1). Given angles α_{ik} and $MDISC_i$ in Tables 3.3, each a_{ik} was calculated by using Equation (6). Given MDL_i the d_i parameters for each item were directly determined by

Table 3.3. Discrimination and Difficulty Levels for the Five Types of Items

Item Type	MDISC	MDL
1	0.4	-1.5
2	0.8	1.0
3	1.2	0.0
4	1.6	-1.0
5	2.0	1.5
Average	1.2	0.0

Equation (6). Since guessing was not considered in this simulation, the M 2-PL compensatory model was used to generate the simulated dichotomous dataset for Study I. Given the MIRT item parameters, the response probability P_{ij} (Equation 2) was computed for each examinee. Then P_{ij} was compared to a uniform random value P^* where $0 \leq P^* \leq 1$. A binary item score of $x_{ij} = 1$ was assigned when $P_{ij} > P^*$. Otherwise, a score of $x_{ij} = 0$ was assigned.

3.1.3 Evaluation Criterion

The criterion variable for evaluating the procedures was the average percentage of items classified into their proper clusters out of the 100 simulation trials. The number of clusters used from the HCA was equal to the number of simulated clusters. For the APSS, the proper clusters also indicated the proper dimensions. In these cases, the lowest possible values were 50% and 33.3% for the simulated two- and three-dimensional tests respectively. For the MS, two different proper clusters were considered. The first proper clusters represented the proper dimensions (i.e., four clusters belong to two dimensions in two-dimensional case). In these cases, the lowest possible average proportions are the same as above (50% and 33.3% for two- and three-dimensional cases). The

second proper clusters represented the proper independent traits and composite traits (four and six clusters for two- and three-dimensional cases). In these cases, the lowest possible values were 25% and 16.7% for the simulated two- and three-dimensional tests respectively.

As Roussos et al. (1998) discussed, these minimum hit rates would result in completely uninterpretable clusters because each cluster would contain an even mixture of items from all the different dimensions or different composite traits. Easily interpretable clusters would require much higher hit rates so that each cluster from the cluster analysis would have the vast majority of its items coming from a single dimension or a single cluster. From examination of the cluster analysis results and recommendations from previous research (Roussos et al., 1998), it seemed that an average hit rate of at least 90% would be needed to result in clusters that are consistently easy to interpret.

In a Monte Carlo examination of 30 different stopping procedures for determining the number of clusters (Milligan and Cooper, 1985), the Calinski and Harabasz (1974) index was evaluated as the best stopping rule to recover the number of clusters. Using this procedure, the hit rates were 88.9%, 88.0%, 89.8% and 94.4% for 2, 3, 4, and 5

clusters. The overall recovery rate was 90.3%. Thus setting a 90% hit rate as a satisfactory level seems reasonable.

For the nonparametric approach, the items in correct clusters were counted from the HCA/CCPROX output. For the parametric approach, the items in correct clusters were counted from the SPSS output (membership grouping) after running HCA with a particular number of clusters specified.

3.2 Study II

The second purpose of this study was to determine if ability levels and guessing affect the results of clustering when using parametric and nonparametric approaches in terms of the proportion of correctly classified items. In this simulation, different levels of ability and guessing were added as variables. To see the effect of different levels of ability and guessing on the clustering results, only the mixed structure conditions were investigated. Thus the variables were somewhat reduced for Study II.

3.2.1 Simulation Factors

The factors of (1) Proximity Measures, (2) HCA Methods, and (3) Number of Examinees were the same as in Study I.

(4) Dimensional Structure

Factors, (a) Number of Dimensions and (b) Number of Items were the same as for Study I.

(c) Number of Clusters: 2 and 4 clusters for 2-D MS tests, and 3 and 6 clusters for 3-D MS tests.

(d) Correlation between Dimensions: Only uncorrelated dimensions were used.

(5) Ability Levels: Lower and higher ability levels.

Lower and higher ability vectors, θ_k, were generated as multivariate normal with means of −.8 for lower ability cases, and .8 for higher ability cases (each has *SD* of 1) on all k dimensions. If a mean of -1.0 is used, then all examinees were observed to answer some difficult items incorrectly. If a mean of +1.0 is used, then all examinees were observed to answer some easiest items correctly. These results were identified through the author's empirical trials. When there is no variation in responses (right and wrong) for an item, then that item cannot be clustered into any group. Thus, -.8 and .8 are limiting values for avoiding all correct or all incorrect response for some items.

(6) Guessing Levels

The value of .17 was used as the guessing level in the study. Since guessing was a factor in this simulation, the M 3-PL compensatory model was used.

3.2.2 Evaluation Criteria

The evaluation criterion for Study II was the difference in the percentages of items correctly classified into the proper dimensions or clusters between Study I and Study II. Of special concern were the cases with satisfactory hit rates (equal to or higher than 90%) in Study I and unsatisfactory hit rates (less than 90%) in Study II.

3.3 Computer Programs

MATLAB 5.3 used to was to generate random numbers, and to compute parametric proximity measures. To apply the nonparametric proximity-based cluster analyses, Roussos's (1992) HCA/CCPROX was used. When the number of examinees is too small (for relatively high or low ability groups), the Pccor is somewhat biased. Thus cases where the number of examinees in a group was less than 20 were eliminated in the computation of Pccor. NOHARM (Fraser, 1988) was used to estimate discrimination parameters for calculating the

angular distance. Among the three kinds of discrimination estimates (unrotated, varimax, and promax) available in NOHARM output, the varimax rotated discrimination estimates were selected for use. This is because the varimax rotation method would tend to align the axis with sets of items. Thus in computing angular distance, use of varimax rotated discrimination estimates make the distinction between clusters more clear. Moreover, the varimax rotated discrimination estimate have the best clustering results in the author's previous empirical trials on a standardized achievement test data. PC-BILOG was used to estimate the *c*-parameters when M 3-P models were used in Study II. SPSS 9.0 was used to implement HCA for the parametric approach.

CHAPTER 4.

RESULTS

Pccor was sometimes not computable for a given item pair because of division by zero. In such cases and the cases when the number of examinees was less than 20 (as stated earlier), CCPROX automatically ignored the easier of the two items and continued the analysis without it. The simulation criterion was calculated only for the non-ignored items. Thus it is important to report the number of items ignored by Pccor. When guessing was considered, HCA/CCPROX did not have any ignored items. Table 4.1 summarizes these results for models without a guessing parameter. Each number indicates the average number of ignored items for Pccor. For example, the value of 8.10 for

Table 4.1. Average Number of Ignored Items for Pccor

Model	Sample size	Average Number of items				
		2-Dimension			3-Dimension	
		20	40	60	30	60
APSS model						
Normal ability	1000	0.00	0.02	0.05	0.00	0.11
	3000	0.00	0.00	0.00	0.00	0.00
MS model						
Normal ability	1000	0.00	0.00	0.00	0.00	0.00
	3000	0.00	0.00	0.00	0.00	0.00
Lower ability	1000	0.30	2.80	8.10	0.20	2.00
	3000	0.00	0.00	0.14	0.00	0.00
Higher ability	1000	0.00	0.00	0.00	0.00	0.00
	3000	0.00	0.00	0.00	0.00	0.00

lower ability level with the 2-D 60 items case indicates that Pccor ignored 8.1 items out of 60 items per each replication. Pccor rarely ignored items except the cases of lower ability without guessing. The lower ability examinee groups were less likely to have enough examinees with same number of correct score.

4.1 Study I

Study I was conducted to compare the effectiveness of HCA methods with parametric and nonparametric approaches to classifying items into clusters. The results are first presented for the approximate simple structure (APSS) and then for the mixed structure (MS) models.

4.1.1 The APSS Models

The results are first presented for the two-dimensional models and then for the three-dimensional models. Table 4.2 shows the average percentages of correctly classified items out of 100 trials for two-dimensional cases. The numbers in the rows labeled Roussos (1998) are from Roussos et al. (1998), so that the results from this study can be compared with the numbers from the earlier study. Most numbers seemed similar.

Table 4.2. Simulation Results for Two-Dimensional APSS

	HCA Methods							
	Single		Complete		Average		Ward	
No. of Items	1000	3000	1000	3000	1000	3000	1000	3000
<20 (10-10) items>								
Non-parametric	55.1	68.2	95.3	98.2	97.6	99.6	95.6	96.3
Roussos(1998)*	58.0	66.7	93.2	98.0	96.9	99.8		
Parametric	98.4	100.0	98.5	100.0	98.3	100.0	98.6	100.0
<40 (20-20) items>								
Non-parametric	52.3	66.8	95.8	98.5	97.3	99.5	94.8	97.3
Roussos(1998)	56.3	79.5	92.7	98.1	98.0	99.5		
Parametric	99.6	100.0	99.7	100.0	99.8	100.0	99.8	100.0
<60 (30-30) items>								
Non-parametric	51.7	90.2	85.8	98.5	96.4	99.6	95.7	97.8
Parametric	99.5	100.0	99.6	100.0	99.7	100.0	99.6	100.0

*The numbers from Roussos et al. (1998)

For the nonparametric approach, all HCA methods seemed to work well except the single link method. The HCA methods worked better with 3000 examinees. Among HCA methods, the average (UPGMA) method gave the best clustering results. With the 3000 sample size, it correctly classified the items into clusters in more than 99% of the trials. For the parametric approach, all HCA methods seemed to work well. All HCA methods were able to classify items correctly into proper clusters more than 98% of the time with 1000 examinees, and to classify perfectly (100%) with 3000 examinees. Like Pccor, all HCA methods with the parametric approach were working better with sample sizes of 3000. The

number of items analyzed did not seem to be related to the hit rates.

Table 4.3 shows the average proportion of correctly classified items out of 100 trials for the three-dimensional cases. Direct comparison between this study and Roussos et al. (1998) was not possible because the methods for randomly generating angles were different.

For the nonparametric approach, single link methods did not work well for either sample size. With 1000 examinees, the complete linkage method did not work well either. Complete, average, and Ward methods seemed to work well with 3000 examinees. Among HCA methods, the average method was able to give excellent results again (99.9%) with the 3000 sample size. For the parametric approach, all HCA methods seemed to work well. All parametric HCA methods were able to classify items correctly into proper clusters more than 91% of the time with 1000 examinees, and to

Table 4.3. Simulation Results for Three-Dimensional APSS

	HCA Methods							
	Single		Complete		Average		Ward	
No. of Items	1000	3000	1000	3000	1000	3000	1000	3000
<30 (10-10-10) items>								
Non-parametric	44.7	47.7	86.7	91.3	95.0	99.9	92.7	94.7
Parametric	91.7	100.0	99.0	100.0	99.7	100.0	99.7	100.0
<60 (20-20-20) items>								
Non-parametric	39.3	67.2	79.8	95.8	94.3	99.3	91.7	97.0
Parametric	92.8	100.0	99.4	100.0	99.3	100.0	99.5	100.0

classify perfectly (100%) with 3000 examinees. Like Pccor, all HCA methods with the parametric approach worked better with larger sample sizes. The number of items (or the number of items attached to a dimension) did not seem to be related to the hit rates in the three-dimensional data.

4.1.2 The MS Model

Table 4.4 shows the average percentages of items correctly classified into their clusters out of 100 trials for two-dimensional cases. Each nonparametric and parametric approach has two rows: 2-C and 4-C. The numbers in each row of 2-C report the average proportions of

Table 4.4. Simulation Results for Two-Dimensional MS

		HCA Methods							
		Single		Complete		Average		Ward	
No. of Items		1000	3000	1000	3000	1000	3000	1000	3000
<20 (10-10) items>									
Non-parametric	2-C	55.0	55.0	95.2	95.5	96.1	96.4	93.7	93.4
	4-C	44.9	45.0	60.0	59.5	59.9	59.2	57.9	61.4
Parametric	2-C	87.1	88.0	92.5	92.9	93.5	93.3	93.7	93.6
	4-C	89.0	89.7	92.4	95.2	91.7	93.9	93.9	96.0
<40 (20-20) items>									
Non-parametric	2-C	52.5	52.5	94.1	98.5	90.8	98.8	89.4	97.2
	4-C	35.0	37.8	59.8	61.3	55.0	54.3	54.8	57.8
Parametric	2-C	81.8	82.2	91.8	94.6	88.7	95.2	91.0	95.2
	4-C	75.0	92.5	92.5	98.8	94.3	99.0	94.0	98.3
<60 (30-30) items>									
Non-parametric	2-C	51.7	51.7	89.5	95.6	96.2	96.6	93.4	94.9
	4-C	28.0	35.8	57.7	58.5	52.8	53.5	56.5	59.5
Parametric	2-C	79.7	80.1	94.8	94.9	95.2	95.6	94.8	94.9
	4-C	77.2	94.7	91.2	99.8	92.3	99.7	94.2	99.7

clustering proper items into two-dimensional clusters, which include dimension 1 (θ_1) clusters (cluster 1 and cluster 3 in (a) of Figure 3.3) and dimension 2 (θ_2) clusters (cluster 2 and cluster 4 in (a) of Figure 3.3). The numbers in the 4-C rows report the average proportions of items properly clustered into 4 clusters, 2 approximate independent item sets (cluster 1 and 2) and 2 composite sets (cluster 3 and 4) in (a) of Figure 3.3.

For the nonparametric approach, HCA methods worked well in classifying items into two dimensions (2-C) at rates higher than 89% with 1000 examinees and at rates higher than 93% with 3000 examinees, except the single link method. The average method seemed to work best, especially with the larger sample size. However, all HCA methods with the nonparametric similarity measures seemed to fail to properly classify items into four clusters. All proportions were less than 62% in these cases.

For the parametric approach, all HCA methods except the single link method worked well to classify items into proper clusters for two dimensions (2-C) at rates higher than 89% with both 1000 and 3000 examinees. Unlike the non-parametric approach, all HCA methods with parametric proximity measures seemed to work well to classify items into 4 clusters. All HCA methods except single link method

showed higher than 90% and 94% of proper classification into 4 clusters, and for the sample size of 3000 even single link method achieved 89% or higher accuracy.

Table 4.5 shows the average proportions of correctly classified items out of 100 trials for three-dimensional cases. Each nonparametric and parametric approach has two rows: the 3-C and 6-C. The numbers in each row of 3-C report the average proportions of correct clustering into three-dimensional clusters, which include dimension 1 (θ_1) clusters (cluster 1 and cluster 4 in (b) of Figure 3.3), dimension 2 (θ_2) clusters (cluster 2 and cluster 5 in (b) of Figure 3.3), and dimension 3 (θ_3) clusters (cluster 3 and cluster 6 in (b) of Figure 3.3). The numbers in each row of 6-C report the average proportions of correct clustering

Table 4.5. Simulation Results for Three-Dimensional MS

		HCA Methods							
		Single		Complete		Average		Ward	
No. of Items		1000	3000	1000	3000	1000	3000	1000	3000
<30 (10-10-10) items>									
Non-parametric	3-C	53.7	51.0	96.8	100.0	99.4	100.0	96.3	98.7
	6-C	39.0	46.7	56.1	57.9	58.3	59.3	53.7	56.7
Parametric	3-C	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	6-C	63.3	67.3	75.0	77.8	73.3	74.7	85.7	86.0
<60 (20-20-20) items>									
Non-parametric	3-C	36.3	35.8	97.0	98.9	99.8	99.9	95.7	96.2
	6-C	24.2	41.1	56.3	56.1	53.5	54.2	51.5	51.7
Parametric	3-C	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	6-C	77.1	99.2	97.9	99.1	98.3	100.0	98.3	99.9

into 6 clusters, which sets consist of three approximate independent dimensions (cluster 1, 2, and 3) and three composite sets (cluster 4, 5, and 6) in (b) of Figure 3.3.

For the nonparametric approach, all HCA methods except the single link method worked well in classifying items into 3 dimensions (3-C) at a hit rate of higher than 95% with both 1000 and 3000 examinees,. The average method seemed to work best with higher than 99% accuracy under both sample sizes. However, for the 6-cluster cases, all HCA methods with nonparametric similarities failed to classify items into proper clusters. All proportions were less than 60%.

For the parametric approach, HCA methods worked well to classify items into 3 dimensions with a perfect hit rate (100%) for both 1000 and 3000 examinees. This result could be partly due to the clear discrepancy (angle gaps) between the 3 dimensional clusters designed in this study. On the other hand, HCA methods with the parametric approach for 6 clusters showed varied hit rates across the number of items. For 30 items all HCA methods did not seem to work well. All classification proportions were equal to or less than 86%, even with 3000 sample size. When test length was 60, all HCA methods except the single link method with both 1000 and 3000 sample sizes showed excellent proportions

(higher than 97%) for clustering into 6 clusters. One reason for the HCA methods' failure for 30 items could be due to the instability of HCA methods to classify small number of items (30) into a relatively large number of clusters (6).

4.2 Study II

Study II investigated the effects of guessing and different ability levels on the clustering efficiency of HCA methods with the parametric and nonparametric approaches. The results are first presented for the guessing effect, then for the ability effect, and finally for the combined ability and guessing effects. Since the single link method has failed to classifying items into proper clusters in Study I, it was excluded from Study II. In addition, since all HCA methods with the nonparametric approach for the 4 and 6 cluster cases (for 2- and 3-Dimensional tests respectively) were unsuccessful, results related to those methods are not explained in the text (the numbers are shown in the tables, however).

4.2.1 Effect of Guessing

Table 4.6 shows the average proportions of items correctly classified out of 100 trials for two-dimensional

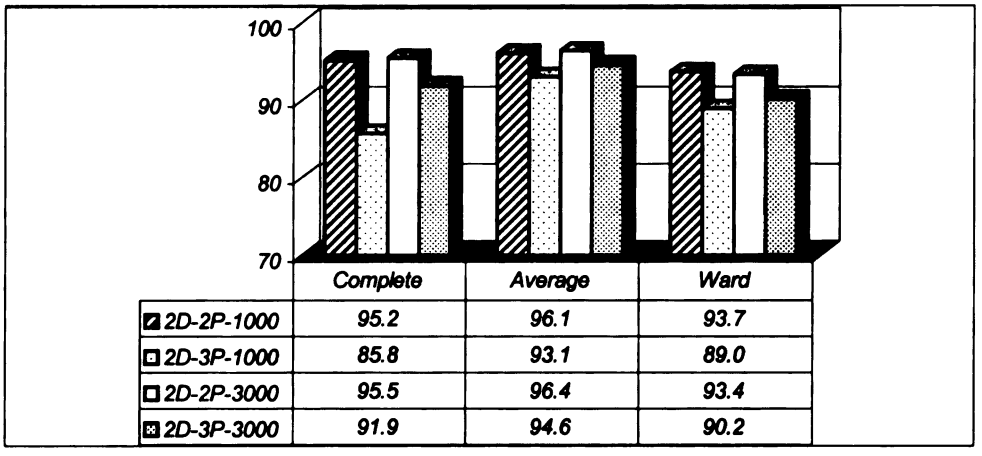
Table 4.6. Simulation Results for Two-Dimensional MS with Guessing

			HCA Methods					
			Complete		Average		Ward	
		Guess	1000	3000	1000	3000	1000	3000
<20 Items>								
	2-C	No	95.2	95.5	96.1	96.4	93.7	93.4
Non-		Yes	85.8	91.9	93.1	94.6	89.0	90.2
Parametric	4-C	No	60.0	59.5	59.9	59.2	57.9	61.4
		Yes	52.4	56.9	56.5	58.2	56.7	55.3
	2-C	No	92.5	92.9	93.5	93.3	93.7	93.6
Parametric		Yes	91.1	91.9	88.6	90.1	91.4	91.6
	4-C	No	92.4	95.2	91.7	93.9	93.9	96.0
		Yes	80.8	92.5	79.9	92.4	85.2	91.4
<40 Items>								
	2-C	No	94.1	98.5	90.8	98.8	89.4	97.2
Non-		Yes	89.7	93.8	90.4	97.0	85.5	96.3
Parametric	4-C	No	59.8	61.3	55.0	54.3	54.8	57.8
		Yes	56.4	58.0	55.6	55.4	56.0	55.8
	2-C	No	91.8	94.6	88.7	95.2	91.0	95.2
Parametric		Yes	91.1	94.6	90.2	95.7	90.8	93.3
	4-C	No	92.5	98.8	94.3	99.0	94.0	98.3
		Yes	87.7	97.2	90.3	97.2	89.3	97.8
<60 Items>								
	2-C	No	89.5	95.6	96.2	96.6	93.4	94.9
Non-		Yes	89.2	93.3	92.7	95.3	82.7	93.1
Parametric	4-C	No	57.7	58.5	52.8	53.5	56.5	59.5
		Yes	54.3	56.1	54.4	53.1	53.8	60.8
	2-C	No	94.8	94.9	95.2	95.6	94.8	94.9
Parametric		Yes	90.3	90.4	90.5	91.1	92.7	92.9
	4-C	No	91.2	99.8	92.3	99.7	94.2	99.7
		Yes	90.5	94.0	92.3	96.4	91.2	95.9

tests without and with guessing. The rows for the 'No' guessing condition were copied from the two-dimension mixed structure results (see Table 4.4).

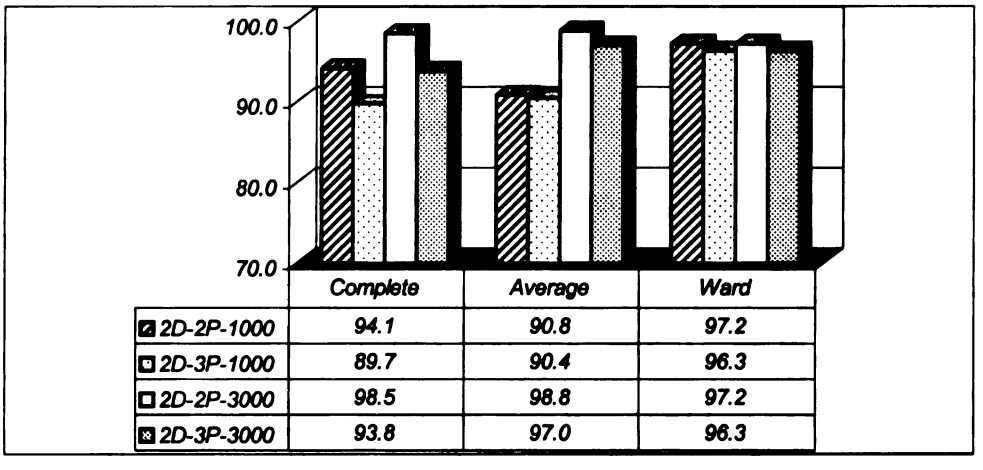
For the nonparametric approach, all three HCA methods showed higher than 90% proper classification into two dimensions (2-C) with 3000 examinees. Among HCA methods, the average method gave the best results. However, proportions for the guessing case were lower than the case without guessing. Figure 4.1 summarizes these results. 2P (2-parameter) indicates no guessing, and 3P (3-parameter) indicates with guessing. Overall there were 17 satisfactory hit rates (equal to or higher than 90%) out of 18 cases (3 HCA methods X 2 sample sizes X 3 test lengths = 18 cases) without guessing. Of these 17, 4 satisfactory hit rates dropped to un-satisfactory rates (lower than 90%) when guessing affected the item responses.

For the parametric approach clustering items into 2 clusters (2-C), all three HCA methods seemed to work well for all test lengths and sample sizes. All three HCA methods had satisfactory hit rates with guessing, except the average method with the sample size 1000 for 20 items. For the parametric approach for 4 clusters (4-C), all three HCA methods seemed to work well for all test lengths especially with the larger sample size. However, the hit



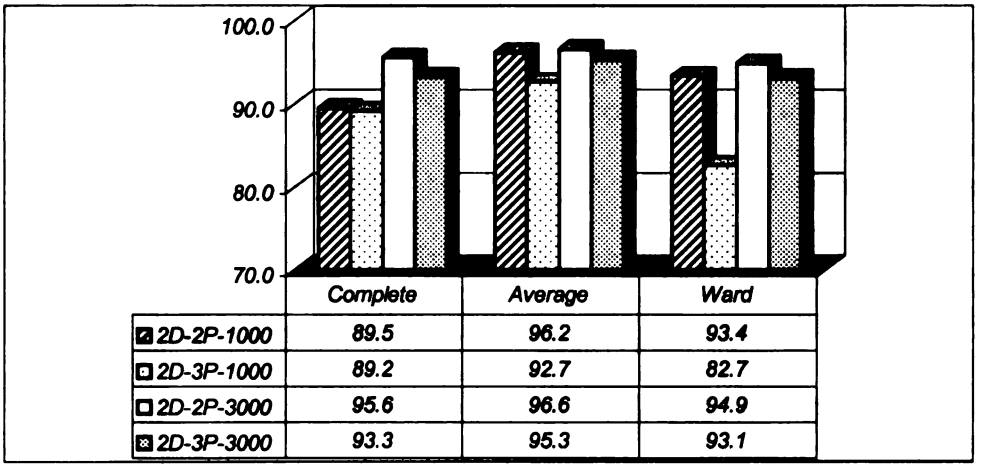
	Complete	Average	Ward
2D-2P-1000	95.2	96.1	93.7
2D-3P-1000	85.8	93.1	89.0
2D-2P-3000	95.5	96.4	93.4
2D-3P-3000	91.9	94.6	90.2

(a) 20 items



	Complete	Average	Ward
2D-2P-1000	94.1	90.8	97.2
2D-3P-1000	89.7	90.4	96.3
2D-2P-3000	98.5	98.8	97.2
2D-3P-3000	93.8	97.0	96.3

(b) 40 Items



	Complete	Average	Ward
2D-2P-1000	89.5	96.2	93.4
2D-3P-1000	89.2	92.7	82.7
2D-2P-3000	95.6	96.6	94.9
2D-3P-3000	93.3	95.3	93.1

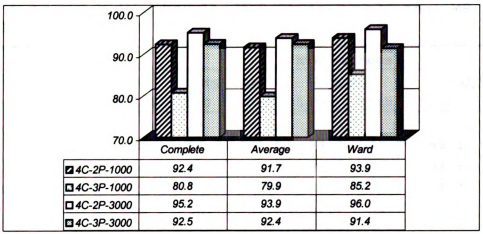
(c) 60 Items

Figure 4.1. Guessing Effect on Correct Classification for 2 Clusters of 2-Dimensional MS: Nonparametric Approach

rates from 'without guessing' to 'with guessing' dropped, especially with the 1000 sample size. With 20 items, the average drop in accuracy was 10.7%. Figure 4.2 shows these results. Overall 5 satisfactory rates dropped to unsatisfactory.

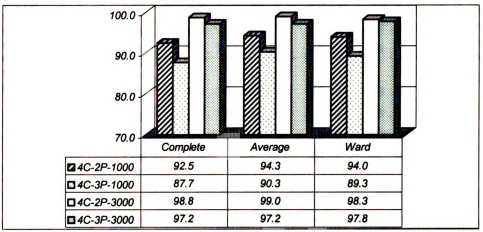
Table 4.7 shows the average proportions of correctly classified items out of 100 trials for three-dimensional mixed structure tests without and with guessing. For the nonparametric approach with guessing, all three HCA methods showed higher than 93% correct classification into three dimensions with both sample sizes. Among HCA methods, the average method gave the best results. However, correct classification with guessing was poorer in comparison to cases without guessing. However, the decline does not seem very large.

For the parametric approach with guessing, all three HCA methods showed higher than 97% correct classification into three dimensions with both 1000 and 3000 examinees. Even though the hit rates dropped when guessing was considered, they were still satisfactory. For classifying items into 6 clusters, HCA methods did not seem to be affected by guessing when 30 was the test length. However, for 60 items, the rates declined quite a bit. These results were due to relatively low initial hit rates for 30 items



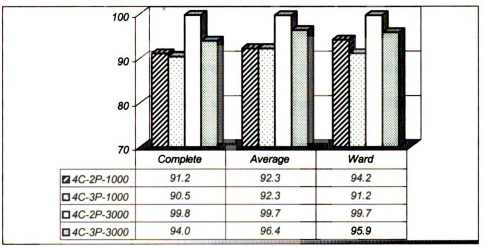
	Complete	Average	Ward
4C-2P-1000	92.4	91.7	93.9
4C-3P-1000	80.8	79.9	85.2
4C-2P-3000	95.2	93.9	96.0
4C-3P-3000	92.5	92.4	91.4

(a) 20 Items



	Complete	Average	Ward
4C-2P-1000	92.5	94.3	94.0
4C-3P-1000	87.7	90.3	89.3
4C-2P-3000	98.8	99.0	98.3
4C-3P-3000	97.2	97.2	97.8

(b) 40 Items



	Complete	Average	Ward
4C-2P-1000	91.2	92.3	94.2
4C-3P-1000	90.5	92.3	91.2
4C-2P-3000	99.8	99.7	99.7
4C-3P-3000	94.0	96.4	95.9

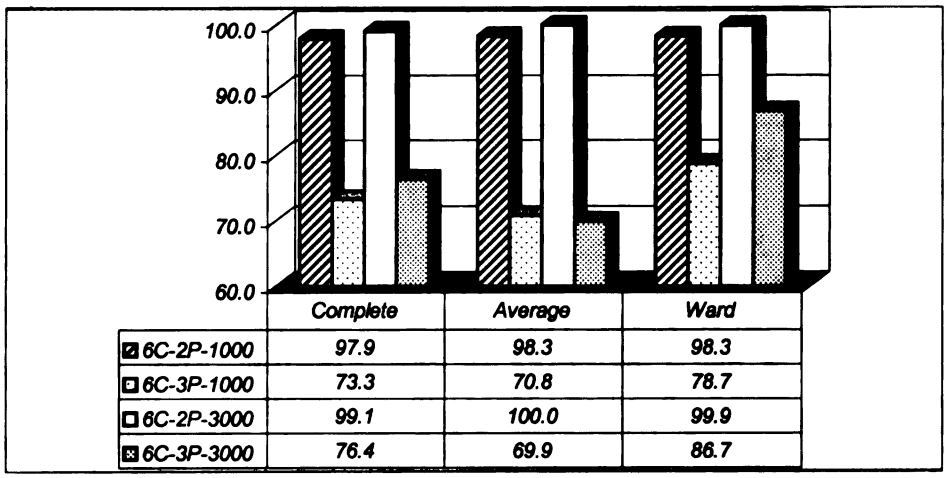
(c) 60 Items

Figure 4.2. Guessing effect on Correct Classification for 4 Clusters of 2-Dimensional MS: Parametric Approach

Table 4.7. Simulation Results for Three-Dimensional MS with Guessing

			HCA Methods					
			Complete		Average		Ward	
		Model	1000	3000	1000	3000	1000	3000
<30 Items>								
	3-C	No	96.8	100.0	99.4	100.0	96.3	98.7
Non-		Yes	95.7	98.3	98.0	99.3	93.3	98.7
Parametric	6-C	No	56.1	57.9	58.3	59.3	53.7	56.7
		Yes	58.7	57.0	57.7	54.7	55.3	57.3
	3-C	No	100.0	100.0	100.0	100.0	100.0	100.0
Parametric		Yes	97.3	99.5	98.3	99.5	98.7	99.7
	6-C	No	75.0	77.8	73.3	74.7	85.7	86.0
		Yes	71.7	77.6	69.3	72.7	76.7	83.0
<60 Items>								
	3-C	No	97.0	98.9	99.8	99.9	95.7	96.2
Non-		Yes	95.0	98.6	96.7	99.5	93.3	98.6
Parametric	6-C	No	56.3	56.1	53.5	54.2	51.5	51.7
		Yes	50.2	55.7	50.7	55.0	51.3	52.1
	3-C	No	100.0	100.0	100.0	100.0	100.0	100.0
Parametric		Yes	99.5	100.0	99.8	100.0	99.7	100.0
	6-C	No	97.9	99.1	98.3	100.0	98.3	99.9
		Yes	73.3	76.4	70.8	69.9	78.7	86.7

(less than 86%) and relatively high initial hit rates for 60 items (higher than 97%) without guessing. The actual rates for 30 and 60 items with guessing were similar to each other. Figure 4.3 shows these big declines for 60 items. All six cases dropped to unsatisfactory rates. The average decline was 20.6%. The average method decreased the most: 27.5% and 30.1% for 1000 and 3000 sample size respectively.



	Complete	Average	Ward
6C-2P-1000	97.9	98.3	98.3
6C-3P-1000	73.3	70.8	78.7
6C-2P-3000	99.1	100.0	99.9
6C-3P-3000	76.4	69.9	86.7

Figure 4.3. Guessing Effect on Correct Classification for 6 Clusters of 3-Dimensional MS: Parametric Approach (60 Items)

4.2.2 Effect of Abilities Levels

Table 4.8 shows the average proportions of correctly classified items out of 100 trials for two-dimensional mixed structure data with lower (mean = -0.8) and higher (mean = 0.8) levels of abilities. The numbers in the row of normal are duplicated from the two-dimensional MS results in Table 4.4.

For the nonparametric approach with both higher and lower ability levels, all three HCA methods yielded higher than 90% correct classification rates for two dimensions with 3000 examinees. However, only the average (UPGMA) method was able to keep the satisfactory hit rates with the 1000 sample size for any test length. Most hit rates with

Table 4.8. Simulation Results for Two-dimensional MS with Different Ability Levels

			HCA Methods					
			Complete		Average		Ward	
		Ability	1000	3000	1000	3000	1000	3000
<20 Items>								
		Low	86.4	93.9	92.2	96.7	85.6	91.1
	2-C	Normal	95.2	95.5	96.1	96.4	93.7	93.4
Non-		High	90.1	95.3	95.0	97.1	90.7	90.8
Parametric								
		Low	55.0	57.5	58.6	59.1	56.4	58.9
	4-C	Normal	60.0	59.5	59.9	59.2	57.9	61.4
		High	57.9	60.4	59.3	58.2	58.4	55.9
		Low	88.1	90.4	88.3	92.2	90.8	92.3
	2-C	Norm	92.5	92.9	93.5	93.3	93.7	93.6
		High	90.1	87.9	88.6	88.6	90.5	91.8
Parametric								
		Low	87.1	94.0	85.8	93.5	89.7	95.2
	4-C	Normal	92.4	95.2	91.7	93.9	93.9	96.0
		High	86.4	94.4	82.9	94.1	88.7	95.5
<40 Items>								
		Low	84.9	91.3	90.1	97.3	90.1	90.3
	2-C	Normal	94.1	98.5	90.8	98.8	89.4	97.2
Non-		High	84.2	94.0	90.4	97.3	88.1	90.4
Parametric								
		Low	56.4	60.0	55.1	55.5	55.3	57.3
	4-C	Normal	59.8	61.3	55.0	54.3	54.8	57.8
		High	58.1	56.4	56.2	54.5	57.4	52.9
		Low	89.7	94.3	88.5	95.2	90.4	94.7
	2-C	Normal	91.8	94.6	88.7	95.2	91.0	95.2
		High	90.3	89.1	90.4	92.3	91.2	92.6
Parametric								
		Low	85.3	86.8	91.1	97.5	90.6	95.5
	4-C	Normal	92.5	98.8	94.3	99.0	94.0	98.3
		High	90.3	97.7	93.8	97.8	93.2	97.3

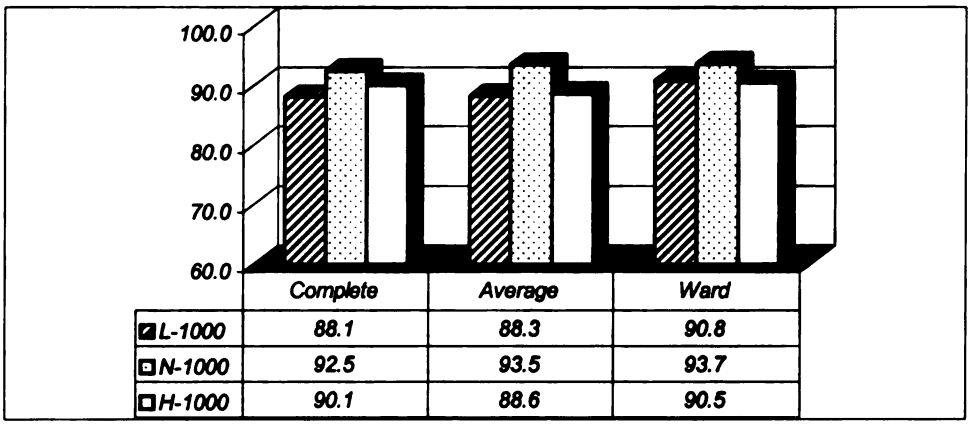
Table 4.8. (Cont'd)

			HCA Methods					
			Complete		Average		Ward	
		Ability	1000	3000	1000	3000	1000	3000
<60 Items>								
		Low	86.3	90.7	95.2	94.5	92.1	90.0
	2-C	Normal	89.5	95.6	96.2	96.6	93.4	94.9
Non-		High	73.2	92.4	92.0	96.1	84.6	95.5
Parametric								
		Low	55.9	54.8	55.1	53.6	54.3	54.5
	4-C	Norm	57.7	58.5	52.8	53.5	56.5	59.5
		High	54.4	56.4	52.3	52.6	54.3	55.2
		Low	89.2	89.8	85.4	89.1	86.9	91.9
	2-C	Normal	94.8	94.9	95.2	95.6	94.8	94.9
		High	85.8	92.6	85.2	89.1	87.1	92.7
Parametric								
		Low	88.9	94.8	91.3	97.4	92.7	96.0
	4-C	Normal	91.2	99.8	92.3	99.7	94.2	99.7
		High	91.7	97.9	92.0	99.1	93.6	99.3

the higher and lower ability levels were reduced from those with average ability level 0.0. Seven satisfactory rates became unsatisfactory. They were for the complete method on lower ability with samples of 1000 for every test length, and the complete method on higher ability with samples of 1000 for 40 and 60 items, for Ward's method on higher abilities with samples of 1000 for 60 items, and Ward's method on lower abilities with samples of 1000 for 20 items.

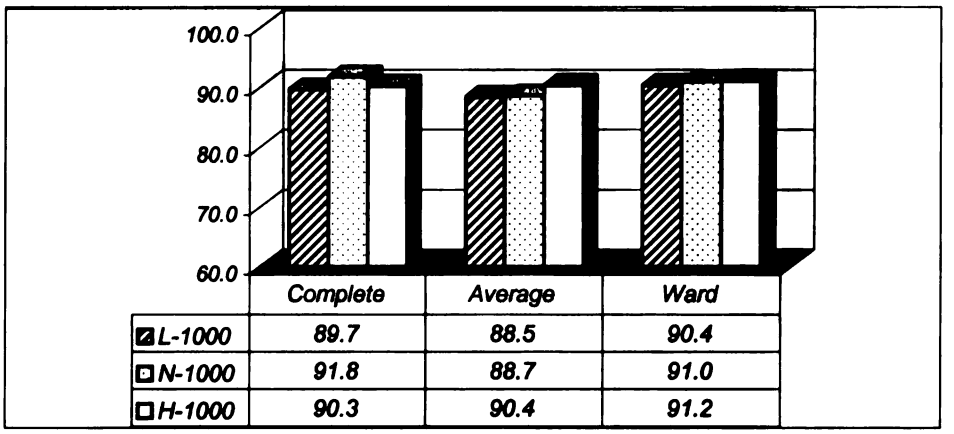
For the parametric approach with both higher and lower ability levels, only Ward's methods with 3000 examinees continued to have higher than 90% correct classifications into 2 clusters. Other HCA methods were affected by different ability levels. Figure 4.4 showed these tendencies for 2-dimensional (2 clusters) MS with the samples of 1000. Four satisfactory hit rates with mean ability 0.0 became unsatisfactory with the higher ability level. Seven satisfactory hit rates with mean ability 0.0 became unsatisfactory with lower ability level. For 60 items, every satisfactory rate became unsatisfactory, though all were higher than 85%.

For the parametric approach with both higher and lower ability levels, average and Ward methods continued to have higher than 90% correct classification into 4 clusters, especially for the 40 and 60 items. With the 1000-examinee samples and 20 items, all HCA methods seemed to be affected by different ability levels. Figure 4.5 shows these tendencies for 4 clusters with the sample of 1000. Three satisfactory rates for ability mean 0.0 with normal became unsatisfactory with the higher ability level (all with 20 items). Five satisfactory rates with ability mean 0.0 became unsatisfactory with lower ability level.



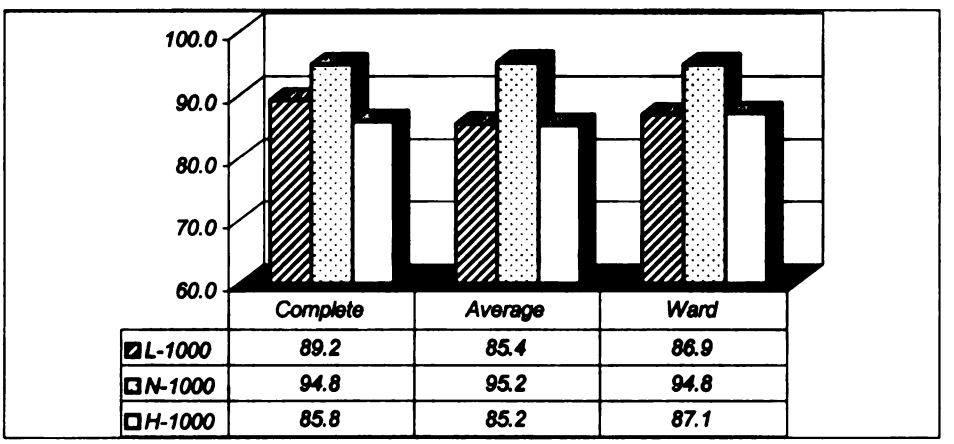
	Complete	Average	Ward
L-1000	88.1	88.3	90.8
N-1000	92.5	93.5	93.7
H-1000	90.1	88.6	90.5

(a) 20 Items



	Complete	Average	Ward
L-1000	89.7	88.5	90.4
N-1000	91.8	88.7	91.0
H-1000	90.3	90.4	91.2

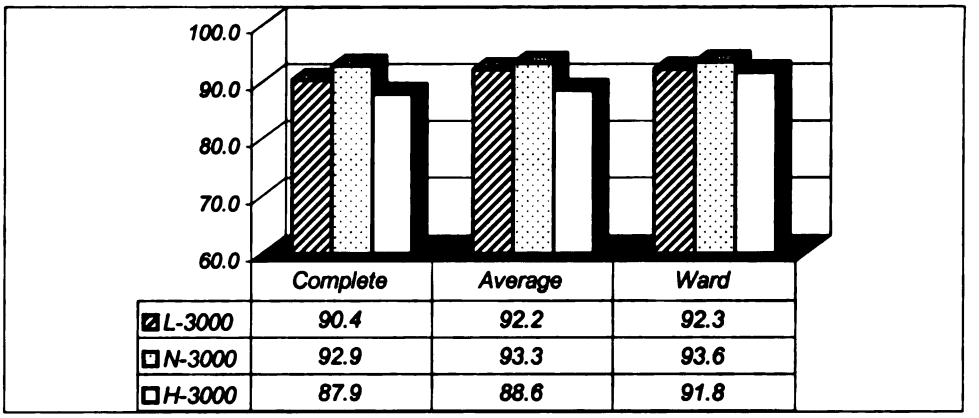
(b) 40 Items



	Complete	Average	Ward
L-1000	89.2	85.4	86.9
N-1000	94.8	95.2	94.8
H-1000	85.8	85.2	87.1

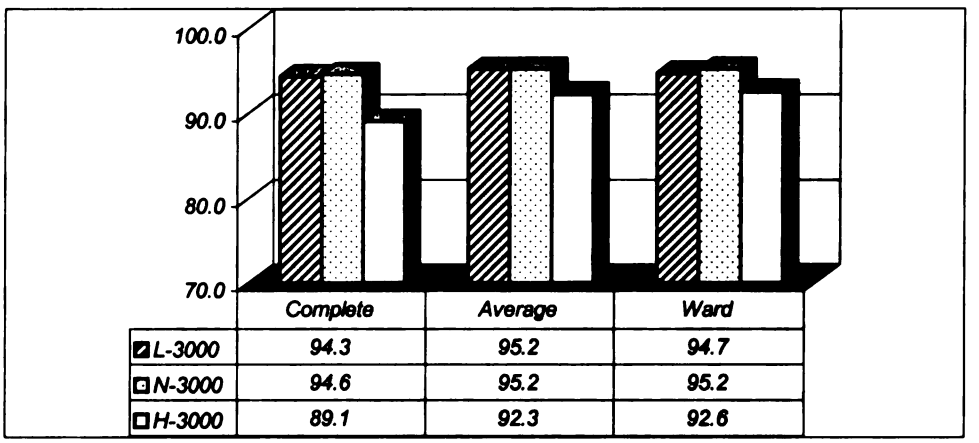
(c) 60 Items

Figure 4.4. Ability Effect on Correct Classification for 2 Clusters of 2-Dimensional MS: Parametric Approach (Sample Size 1000)



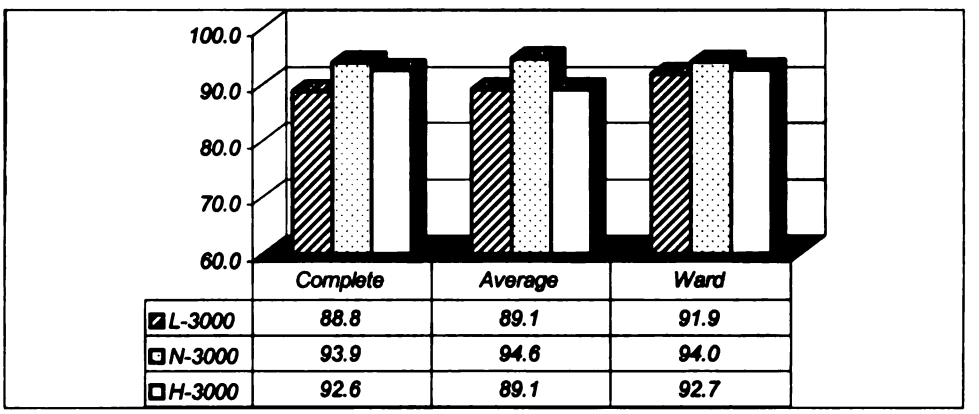
	Complete	Average	Ward
L-3000	90.4	92.2	92.3
N-3000	92.9	93.3	93.6
H-3000	87.9	88.6	91.8

(a) 20 Items



	Complete	Average	Ward
L-3000	94.3	95.2	94.7
N-3000	94.6	95.2	95.2
H-3000	89.1	92.3	92.6

(b) 40 Items



	Complete	Average	Ward
L-3000	88.8	89.1	91.9
N-3000	93.9	94.6	94.0
H-3000	92.6	89.1	92.7

(c) 60 Items

Figure 4.5. Ability Effect on Correct Classification for 4 Clusters with 2-Dimensional MS: Parametric Approach (Sample Size 1000)

Table 4.9 shows the average proportions of correctly classified items out of 100 trials for three-dimensional mixed structure with the lower and higher levels of abilities. For the nonparametric approach with both higher and lower ability levels, all three HCA methods showed higher than 92% correct classifications for three dimensions with both 1000 and 3000 examinees regardless of the number of items. Among HCA methods, the average method gave the best results.

For the parametric approach with both higher and lower ability levels, all three HCA methods showed higher than 96% correct classification into three dimensions, regardless of the number of items and number of examinees. It was remarkable that all HCA methods recovered 100% correct classification when sample sizes were 3000. The different abilities did not affect the hit rates of all HCA methods for classification into six clusters. For 60 items, all HCA methods were satisfactory with higher than 94% correct classifications with two different ability levels.

4.2.3 Effects of the Combination of Different Abilities and Guessing

Table 4.10 shows the average proportions of correctly classified items into their dimensions out of 100 trials

Table 4.9. Simulation Results for Three-Dimensional MS with Different Ability Levels

			HCA Methods					
			Complete		Average		Ward	
		Ability	1000	3000	1000	3000	1000	3000
<30 Items>								
		Lower	92.0	99.1	98.3	99.8	93.3	99.1
	3-C	Normal	96.8	100.0	99.4	100.0	96.3	98.7
Non-		Higher	97.1	98.7	99.3	99.0	95.7	96.0
Parametric								
		Lower	59.3	58.9	58.0	57.7	56.3	53.3
	6-C	Normal	56.1	57.9	58.3	59.3	53.7	56.7
		Higher	56.9	57.3	58.2	59.7	53.7	52.0
		Lower	99.1	100.0	99.0	100.0	99.3	100.0
	3-C	Normal	100.0	100.0	100.0	100.0	100.0	100.0
		Higher	100.0	100.0	100.0	100.0	100.0	100.0
Parametric								
		Lower	74.8	77.8	67.7	74.7	79.3	85.6
	6-C	Normal	75.0	77.8	73.3	74.7	85.7	86.0
		Higher	75.3	78.5	67.7	76.7	79.3	85.3
<60 Items>								
		Lower	96.9	98.1	98.4	99.8	94.7	95.1
	3-C	Normal	97.0	98.9	99.8	99.9	95.7	96.2
Non-		Higher	92.2	99.1	98.2	99.7	96.0	96.2
Parametric								
		Lower	54.7	58.1	53.2	54.6	51.9	53.6
	6-C	Normal	56.3	56.1	53.5	54.2	51.5	51.7
		Higher	59.2	57.6	54.3	54.5	55.3	52.4
		Lower	97.8	100.0	96.9	100.0	98.2	100.0
	3-C	Normal	100.0	100.0	100.0	100.0	100.0	100.0
		Higher	100.0	100.0	100.0	100.0	100.0	100.0
Parametric								
		Lower	94.8	99.3	94.7	99.4	96.6	99.5
	6-C	Normal	97.9	99.1	98.3	100.0	98.3	99.9
		Higher	97.2	99.7	98.2	99.5	98.5	99.6

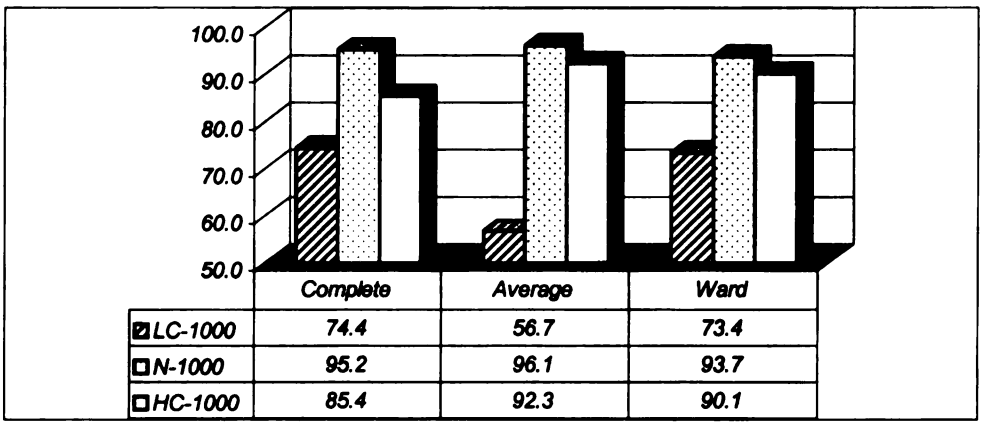
Table 4.10. Simulation Results for Two-Dimensional MS with Combination of Different Abilities and Guessing

			HCA Methods					
			Complete		Average		Ward	
			1000	3000	1000	3000	1000	3000
<20 items>								
		Low a & c	74.4	81.1	56.7	58.4	73.4	80.0
	2-C	Normal	95.2	95.5	96.1	96.4	93.7	93.4
Non-		High a & c	85.4	93.6	92.3	95.0	90.1	92.3
Parametric								
		Low a & c	49.9	53.3	49.4	47.3	51.6	54.2
	4-C	Normal	60.0	59.5	59.9	59.2	57.9	61.4
		High a & c	54.9	57.4	58.1	58.9	57.2	56.5
		Lower a & c	78.8	86.4	78.5	84.3	84.4	80.9
	2-C	Normal	92.5	92.9	93.5	93.3	93.7	93.6
		High a & c	90.8	90.7	88.7	92.2	91.1	91.9
Parametric								
		Lowe a & c	70.5	77.8	70.4	76.5	71.3	80.1
	4-C	Normal	92.4	95.2	91.7	93.9	93.9	96.0
		High a & c	79.6	93.9	79.5	94.1	83.4	92.5
<40 items>								
		Low a & c	80.1	89.8	80.2	92.5	67.8	77.3
	2-C	Normal	94.1	98.5	90.8	98.8	89.4	97.2
Non-		High a & c	85.2	92.8	90.9	96.5	86.2	86.9
Parametric								
		Low a & c	49.7	52.8	51.5	52.8	49.2	51.0
	4-C	Normal	59.8	61.3	55.0	54.3	54.8	57.8
		High a & c	55.6	57.0	56.2	53.7	56.7	55.3
		Low a & c	80.6	92.2	64.7	94.6	86.7	90.3
	2-C	Normal	91.8	94.6	88.7	95.2	91.0	95.2
		High a & c	66.8	90.8	60.7	97.1	75.9	94.9
Parametric								
		Low a & c	83.0	83.8	83.3	76.4	87.1	90.5
	4-C	Normal	92.5	98.8	94.3	99.0	94.0	98.3
		High a & c	89.7	93.2	89.8	93.9	89.3	94.2

Table 4.10. (Cont'd)

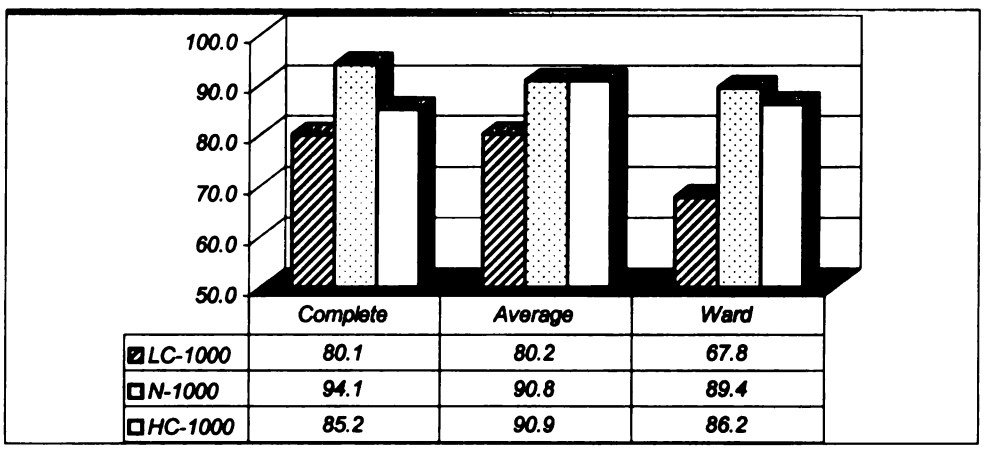
			HCA Methods					
			Complete		Average		Ward	
			1000	3000	1000	3000	1000	3000
<60 items>								
		Low a & c	75.8	84.1	83.7	92.4	68.5	81.7
	2-C	Normal	89.5	95.6	96.2	96.6	93.4	94.9
Non-		High a & c	68.3	90.2	93.4	92.4	89.5	89.8
Parametric								
		Low a & c	51.8	49.5	52.8	52.4	52.7	50.7
	4-C	Normal	57.7	58.5	52.8	53.5	56.5	59.5
		High a & c	55.1	56.4	55.9	53.6	56.6	52.9
		Low a & c	82.2	82.8	70.2	85.2	81.5	91.7
	2-C	Normal	94.8	94.9	95.2	95.6	94.8	94.9
		High a & c	89.2	86.9	93.8	96.7	89.8	96.8
Parametric								
		Low a & c	61.1	83.8	63.7	76.4	73.7	90.5
	4-C	Normal	91.2	99.8	92.3	99.7	94.2	99.7
		High a & c	85.3	95.1	87.3	95.2	90.7	95.5

for two-dimensional cases with the combined lower ability and guessing (low a & c), and with the combined higher ability and guessing (high a & c). For the nonparametric approach, all three HCA methods seemed to be affected by the combination of different abilities and guessing. Only the average method for 40 and 60 items had higher than 90% correct classification into two dimensions with 3000 examinees. Figure 4.6 shows the combined effect of different abilities and guessing on the proportion of classification for the 1000 sample size. Three that were



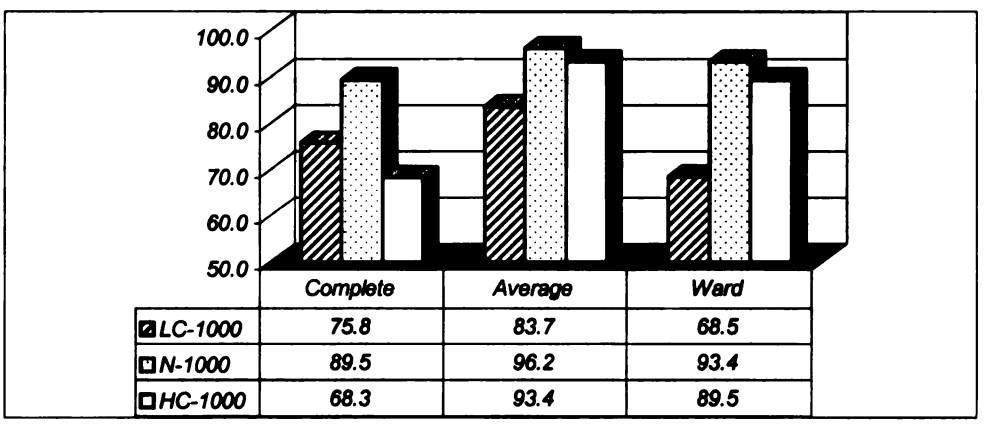
	Complete	Average	Ward
LC-1000	74.4	56.7	73.4
N-1000	95.2	96.1	93.7
HC-1000	85.4	92.3	90.1

(a) 20 Items



	Complete	Average	Ward
LC-1000	80.1	80.2	67.8
N-1000	94.1	90.8	89.4
HC-1000	85.2	90.9	86.2

(b) 40 Items



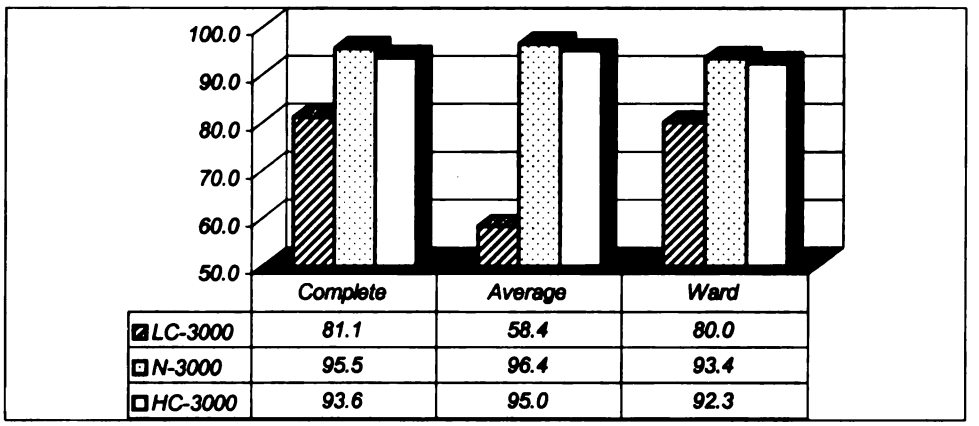
	Complete	Average	Ward
LC-1000	75.8	83.7	68.5
N-1000	89.5	96.2	93.4
HC-1000	68.3	93.4	89.5

(c) 60 items

Figure 4.6. Abilities and Guessing Combined Effect on Correct Classification for 2 Clusters of 2-Dimensional MS: Nonparametric Approach (Sample Size 1000)

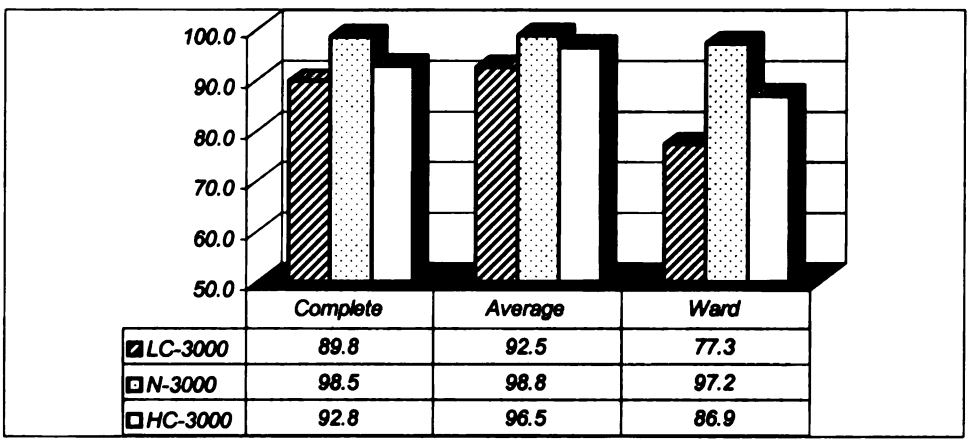
satisfactory for the mean ability 0.0 data became unsatisfactory with the combination of higher ability and guessing. For the combined higher-ability and guessing, the average drop in rates were 5.7%, 4.3%, and 9.3% for 20, 40, and 60 items respectively. The complete linkage method showed the largest drop in rate. All 7 satisfactory hit rates with mean 0.0 data became unsatisfactory with the combination of lower ability and guessing. For the combined lower-ability and guessing, the average drop rates were 26.8%, 15.4%, and 17.0% for 20, 40, and 60 items respectively.

Figure 4.7 shows the effect of the combination of different abilities and guessing levels on the proportions of correct classification for sample size of 3000. Only two satisfactory hit rates (with mean 0.0 data) became unsatisfactory with the combined higher ability and guessing. Seven satisfactory hit rates with mean ability of 0.0 became unsatisfactory with the combination of lower ability and guessing. For lower ability and guessing combined, the average drop in rates were 21.9%, 11.6%, and 9.6% for 20, 40, and 60 items respectively. Overall all HCA methods using nonparametric similarities seemed to be affected by the combination of ability and guessing, particularly for low ability distributions.



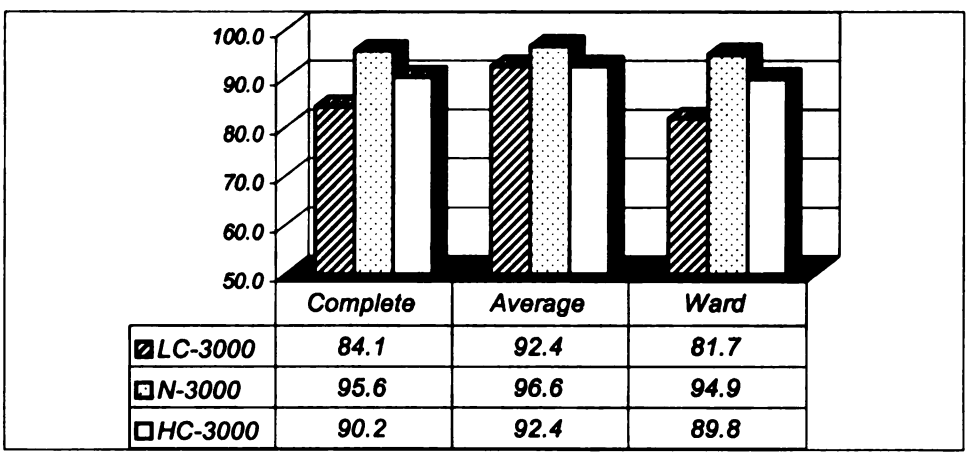
	Complete	Average	Ward
LC-3000	81.1	58.4	80.0
N-3000	95.5	96.4	93.4
HC-3000	93.6	95.0	92.3

(a) 20 items



	Complete	Average	Ward
LC-3000	89.8	92.5	77.3
N-3000	98.5	98.8	97.2
HC-3000	92.8	96.5	86.9

(b) 40 items



	Complete	Average	Ward
LC-3000	84.1	92.4	81.7
N-3000	95.6	96.6	94.9
HC-3000	90.2	92.4	89.8

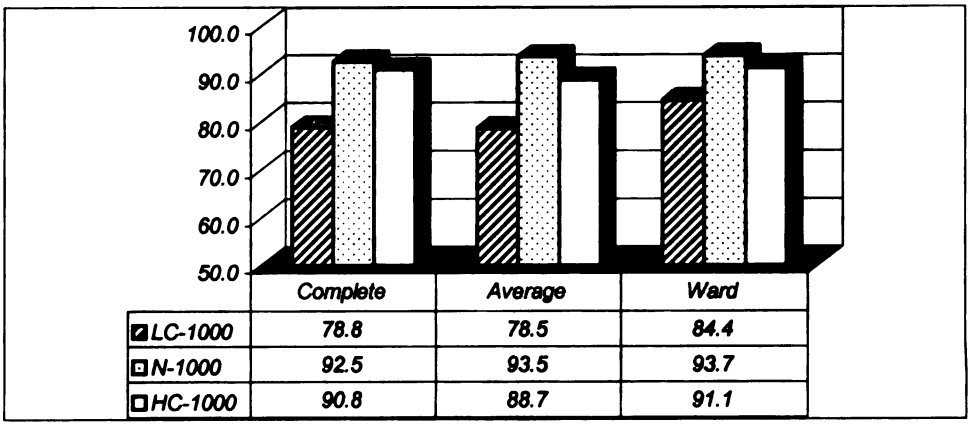
(c) 60 Items

Figure 4.7. Abilities and Guessing Combined Effect on Correct Classification for 2 Clusters of 2-Dimensional MS: Nonparametric Approach (Sample Size 3000)

For the parametric approach, all HCA methods were somewhat influenced by the combination of different abilities and guessing level. However, most HCA methods for 40 and 60 items continued to have higher than 90% correct classification into two dimensions with 3000 examinees for the combination of higher ability and guessing.

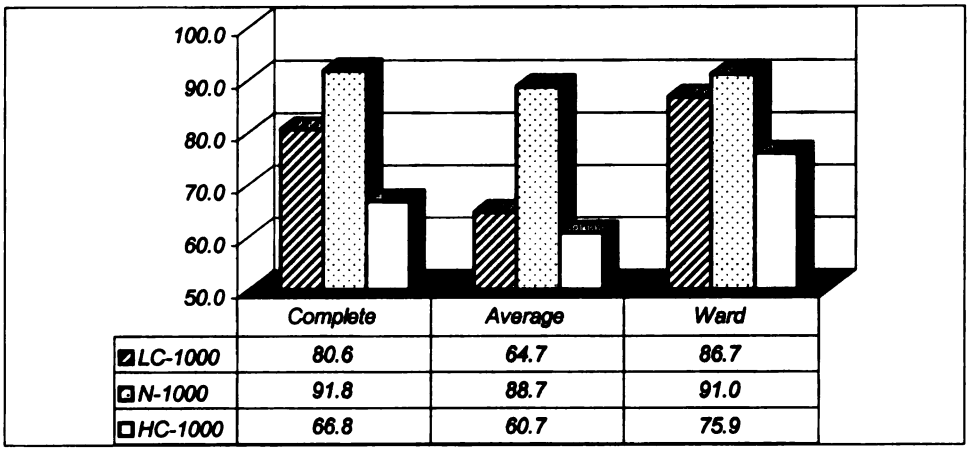
Figure 4.8 shows the effect of the combination of different abilities and guessing on the proportion of correct classification for sample sizes of 1000. Five satisfactory hit rates with mean 0.0 ability data became unsatisfactory with the combination of higher ability and guessing. Eight satisfactory hit rates with mean 0.0 data became unsatisfactory with the combination of lower ability and guessing. For the combination of lower ability and guessing, the average drops were 12.7%, 13.2%, and 20.3% for 20, 40, and 60 items respectively. For the combination of higher ability and guessing, the average drop in rates for 40 items was 22.7%.

Figure 4.9 shows the effects of different abilities and guessing levels on the proportion of correct classification for the 3000 sample size. The higher ability with guessing did not seem to affect HCA methods' classification rates (except the complete linkage method for 60 items). However, all HCA methods for 20 items, and



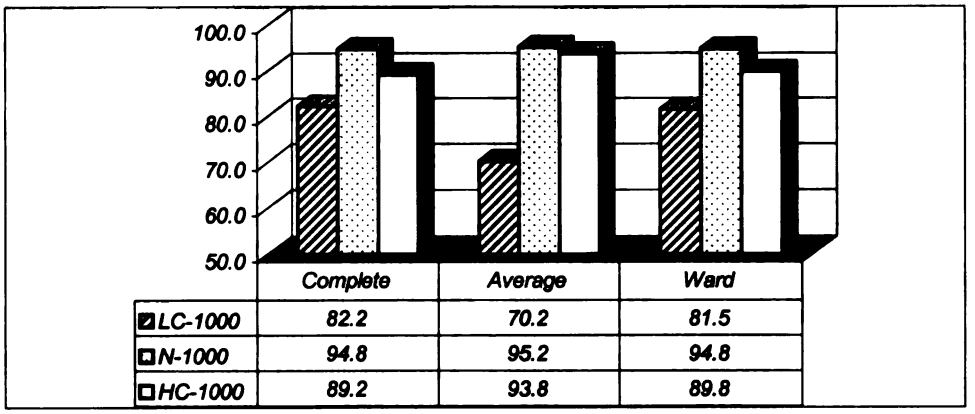
	Complete	Average	Ward
LC-1000	78.8	78.5	84.4
N-1000	92.5	93.5	93.7
HC-1000	90.8	88.7	91.1

(a) 20 Items



	Complete	Average	Ward
LC-1000	80.6	64.7	86.7
N-1000	91.8	88.7	91.0
HC-1000	66.8	60.7	75.9

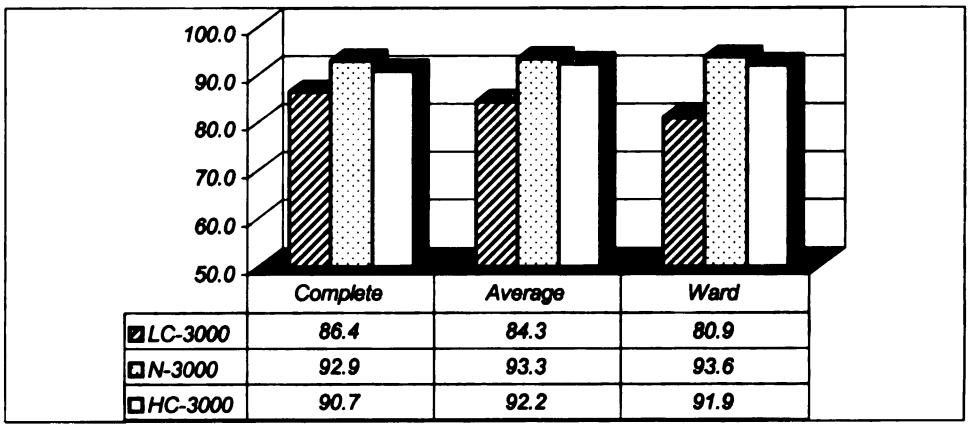
(b) 40 Items



	Complete	Average	Ward
LC-1000	82.2	70.2	81.5
N-1000	94.8	95.2	94.8
HC-1000	89.2	93.8	89.8

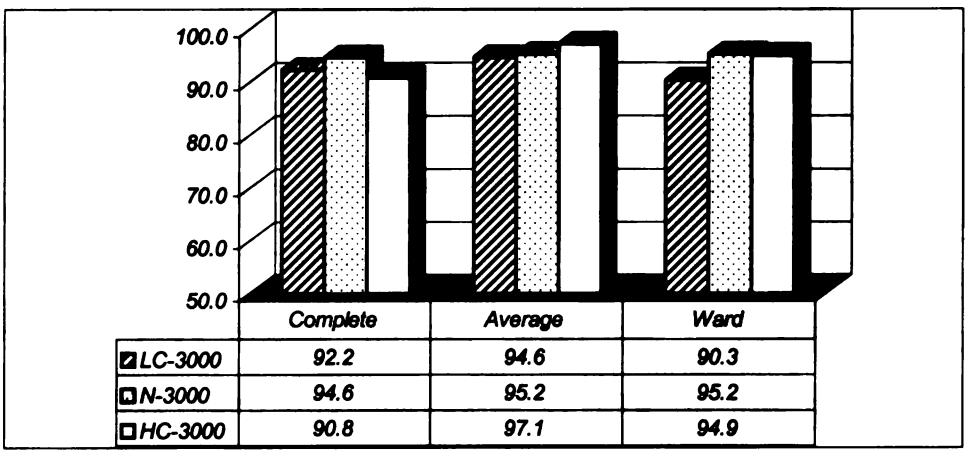
(c) 60 items

Figure 4.8. Abilities and Guessing Combined Effect on Correct Classification for 2 Clusters of 2-Dimensional MS: Parametric Approach (Sample Size 1000)



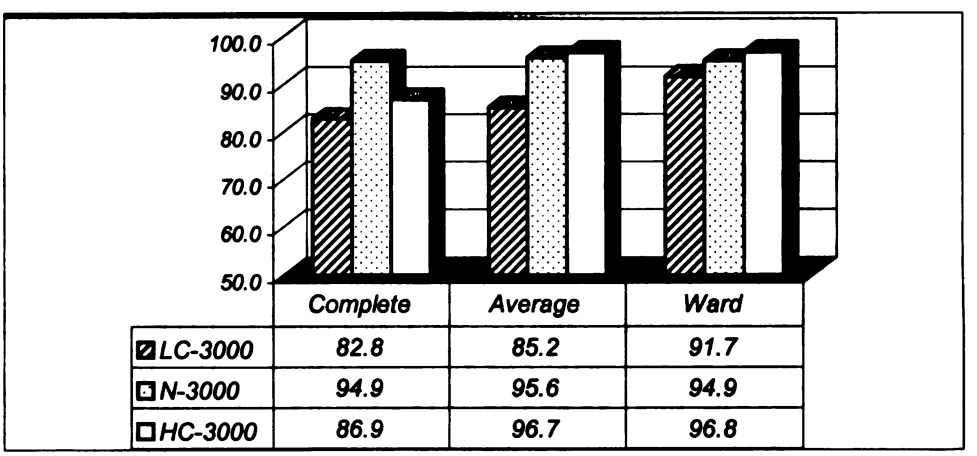
	Complete	Average	Ward
LC-3000	86.4	84.3	80.9
N-3000	92.9	93.3	93.6
HC-3000	90.7	92.2	91.9

(a) 20 Items



	Complete	Average	Ward
LC-3000	92.2	94.6	90.3
N-3000	94.6	95.2	95.2
HC-3000	90.8	97.1	94.9

(b) 40 Items



	Complete	Average	Ward
LC-3000	82.8	85.2	91.7
N-3000	94.9	95.6	94.9
HC-3000	86.9	96.7	96.8

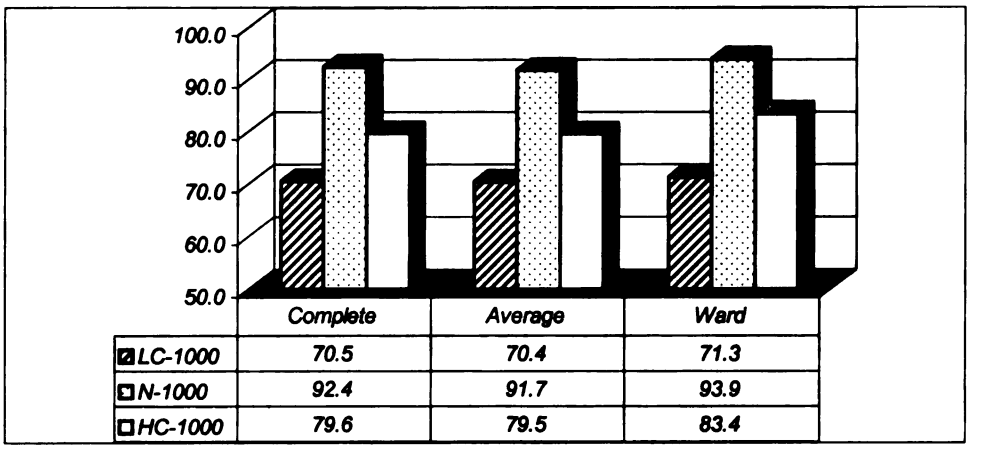
(c) 60 Items

Figure 4.9. Abilities and Guessing Combined Effect on Correct Classification for 2 Clusters of 2-Dimensional MS: Parametric Approach (Sample Size 3000)

the complete linkage and average method for 60 items were affected by the combination of lower ability and guessing. The average decline rates for 20 and 60 items were 8.5% and 8.6% respectively.

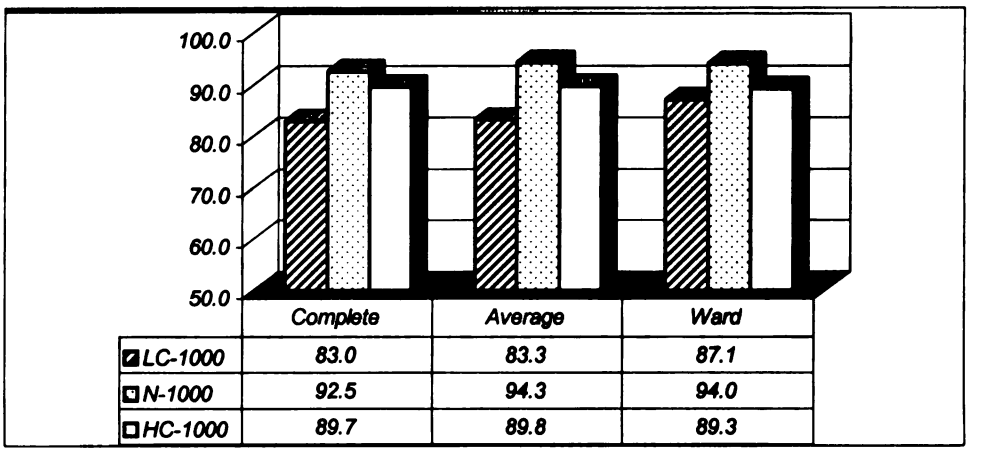
For the parametric approach applied to four cluster data, all three HCA methods were somewhat affected by the combination of different abilities and guessing levels. However, regardless of test length all HCA methods classified items at a higher than 90% rate for four clusters with 3000 examinees under the combination of higher ability and guessing.

Figure 4.10 shows the effect of combination of different abilities and guessing levels on the proportion of correct classification for the sample sizes of 1000. All satisfactory hit rates with mean ability 0.0 became unsatisfactory with the combined higher ability level and guessing (except the Ward's method with 60 items). All nine satisfactory hit rates for mean ability 0.0 became un-satisfactory with the combined lower ability and guessing. For the combination of higher ability and guessing, the average rates dropped 11.8%, 4.0%, and 4.8% for 20, 40, and 60 items, respectively. For the combination of lower ability and guessing, the average drops were 21.9%, 9.1%, and 23.1% for 20, 40, and 60 items respectively.



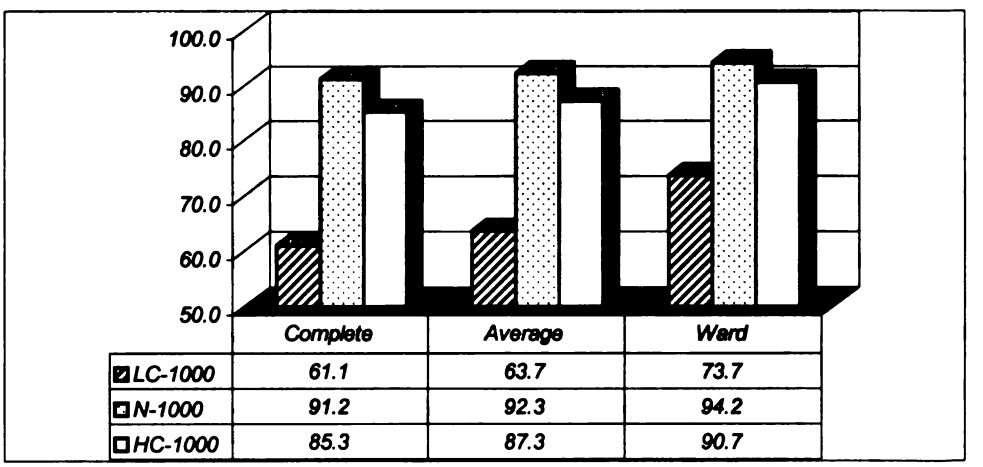
	Complete	Average	Ward
LC-1000	70.5	70.4	71.3
N-1000	92.4	91.7	93.9
HC-1000	79.6	79.5	83.4

(a) 20 Items



	Complete	Average	Ward
LC-1000	83.0	83.3	87.1
N-1000	92.5	94.3	94.0
HC-1000	89.7	89.8	89.3

(b) 40 items



	Complete	Average	Ward
LC-1000	61.1	63.7	73.7
N-1000	91.2	92.3	94.2
HC-1000	85.3	87.3	90.7

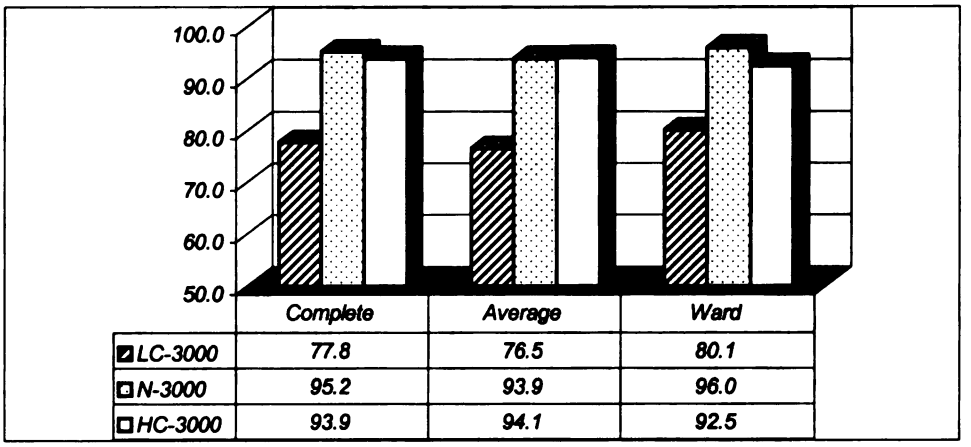
(c) 60 Items

Figure 4.10. Ability and Guessing Combined Effect on Correct Classification for 4 Clusters of 2-Dimensional MS: Parametric Approach (Sample Size 1000)

Figure 4.11 shows the combined effect of different abilities and guessing levels on the proportion of correct classification for four clusters and 3000 cases. Higher ability with guessing did not seem to affect HCA methods' classification rates even though some proper classification rates declined. However, most HCA methods seemed to be affected by the combination of lower ability and guessing (seven satisfactory hit rates became unsatisfactory). Average declines in hit rates were 16.9%, 15.1% and 16.2% for 20, 40, and 60 items respectively.

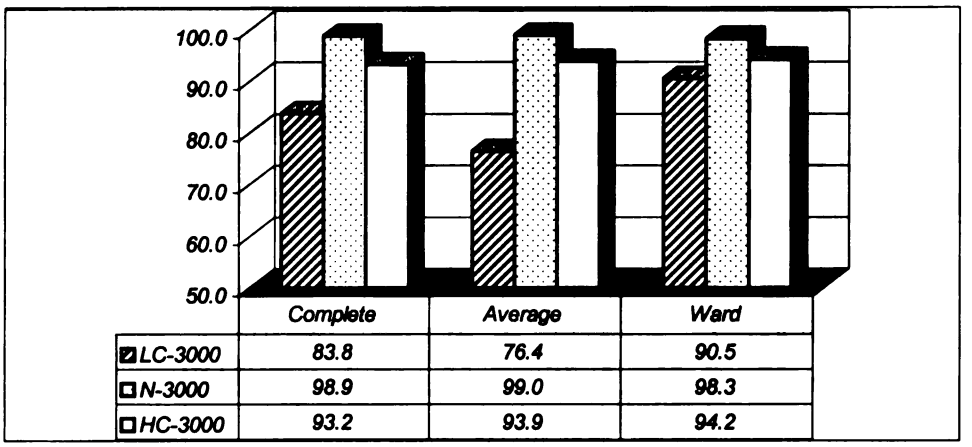
Table 4.11 shows the average proportions of correctly classified items out of 100 trials for three-dimensional cases with lower and higher levels of abilities with guessing.

For the nonparametric approach, all three HCA methods did not seem to be affected by the combination of high ability and guessing, but were affected by the combination of lower ability and guessing situation. Only the average method for 60 items had higher than 90% correct classifications into three dimensions with 3000 examinees with the combination of lower ability and guessing. Figure 4.12 shows the effect of the combination of different abilities and guessing on the correct classification proportions for 3 dimensions for the nonparametric approach. The average



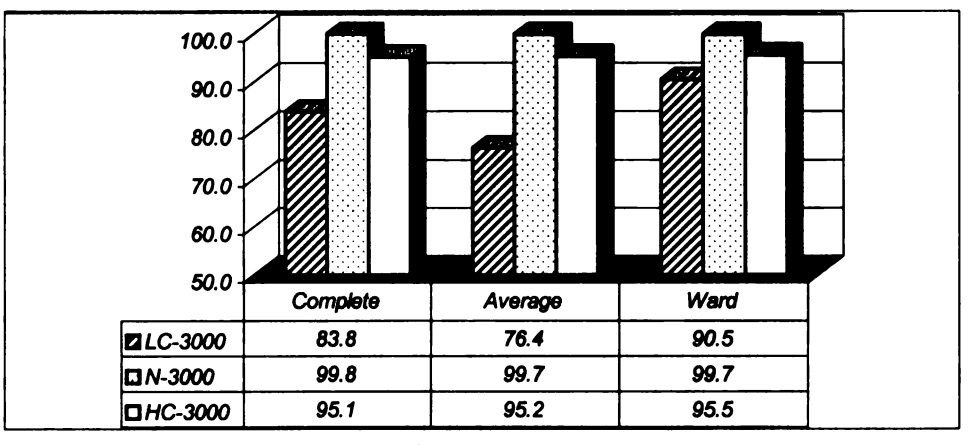
	Complete	Average	Ward
LC-3000	77.8	76.5	80.1
N-3000	95.2	93.9	96.0
HC-3000	93.9	94.1	92.5

(a) 20 Items



	Complete	Average	Ward
LC-3000	83.8	76.4	90.5
N-3000	98.9	99.0	98.3
HC-3000	93.2	93.9	94.2

(b) 40 Items



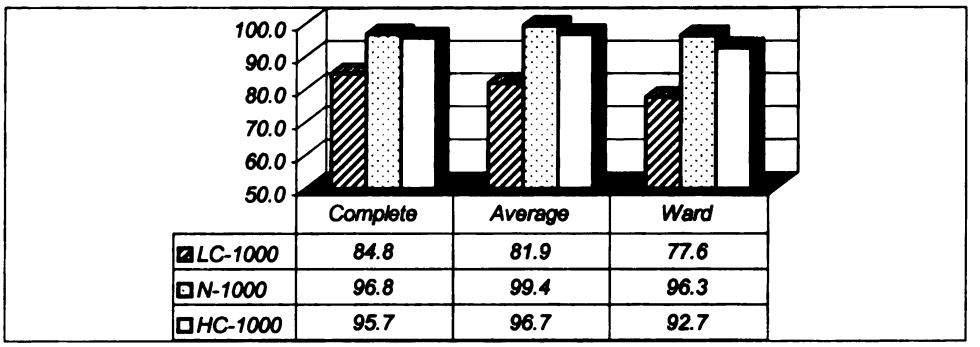
	Complete	Average	Ward
LC-3000	83.8	76.4	90.5
N-3000	99.8	99.7	99.7
HC-3000	95.1	95.2	95.5

(c) 60 items

Figure 4.11. Ability and Guessing Combined Effect on Correct Classification for 4 Clusters of 2-Dimensional MS: Parametric Approach (Sample Size 3000)

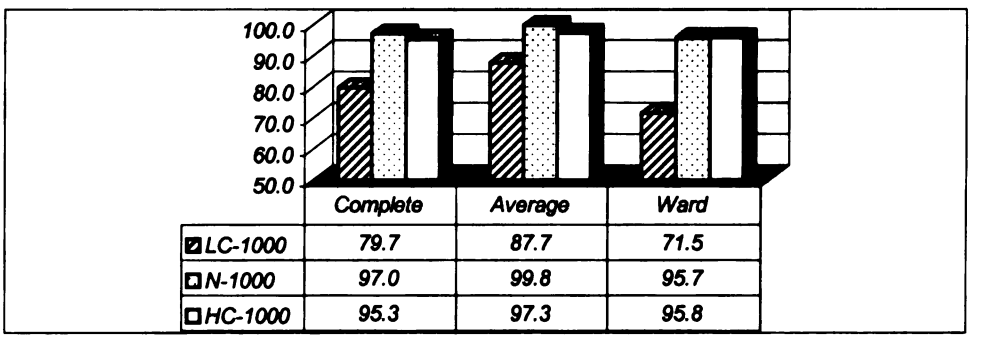
Table 4.11. Simulation Results for Three-Dimensional MS with Combination of Different Abilities and Guessing

			HCA Methods					
			Complete		Average		Ward	
			1000	3000	1000	3000	1000	3000
<30 Items>								
		Lower a & c	84.8	89.0	81.9	82.7	77.6	76.9
	3-C	Normal	96.8	100.0	99.4	100.0	96.3	98.7
Non-		Higher a & c	95.7	99.3	96.7	99.7	92.7	96.3
Parametric								
		Lower a & c	53.9	55.5	55.3	54.7	52.0	50.7
	6-C	Normal	56.3	57.9	58.3	59.3	53.7	56.7
		Higher a & c	56.7	57.3	57.1	58.7	55.7	54.0
		Lower a & c	91.0	94.3	75.7	89.3	93.1	94.3
	3-C	Normal	100.0	100.0	100.0	100.0	100.0	100.0
		Higher a & c	98.7	100.0	99.2	100.0	99.3	100.0
Parametric								
		Lower a & c	60.4	68.1	61.7	61.9	62.4	65.7
	6-C	Normal	75.0	77.8	73.3	74.7	85.7	86.0
		Higher a & c	72.7	77.2	65.7	71.0	76.7	81.4
<60 Items>								
		Lower a & c	79.7	89.3	87.7	94.0	71.5	76.1
	3-C	Normal	97.0	98.9	99.8	99.9	95.7	96.2
Non-		Higher a & c	95.3	98.1	97.3	99.5	95.8	94.3
Parametric								
		Lower a & c	43.8	47.9	47.2	52.1	43.9	51.2
	6-C	Normal	56.5	56.1	53.5	54.2	51.5	51.7
		Higher a & c	52.8	52.8	52.4	54.1	51.4	51.2
		Lower a & c	77.3	84.3	78.5	81.3	92.3	86.3
	3-C	Normal	100.0	100.0	100.0	100.0	100.0	100.0
		Higher a & c	99.3	100.0	99.6	100.0	99.8	100.0
Parametric								
		Lower a & c	57.1	65.3	51.7	59.7	60.8	74.5
	6-C	Normal	97.9	99.1	98.3	100.0	98.3	99.9
		Higher a & c	77.2	82.4	68.3	68.8	72.6	88.8



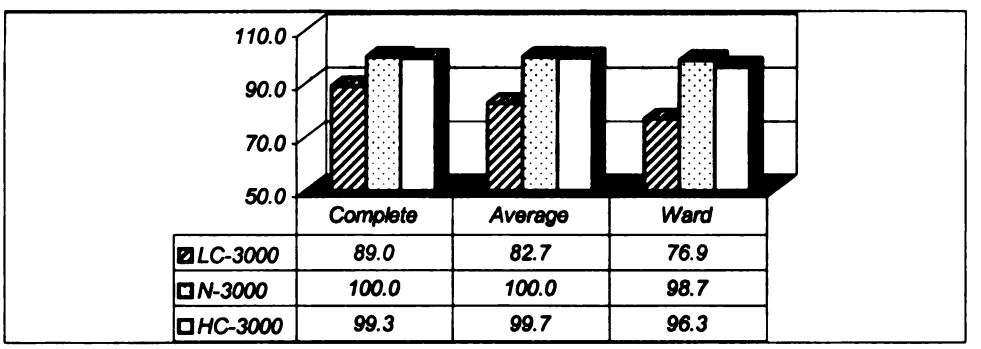
	Complete	Average	Ward
LC-1000	84.8	81.9	77.6
N-1000	96.8	99.4	96.3
HC-1000	95.7	96.7	92.7

(a) 30 Items with sample size 1000



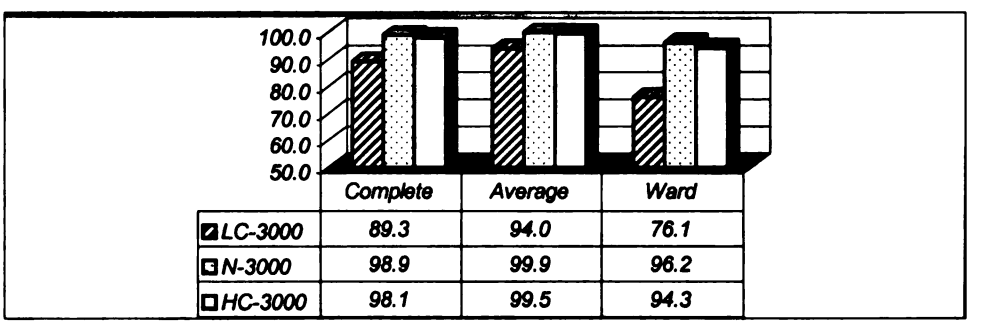
	Complete	Average	Ward
LC-1000	79.7	87.7	71.5
N-1000	97.0	99.8	95.7
HC-1000	95.3	97.3	95.8

(b) 60 Items with sample size 1000



	Complete	Average	Ward
LC-3000	89.0	82.7	76.9
N-3000	100.0	100.0	98.7
HC-3000	99.3	99.7	96.3

(c) 30 Items with sample size 3000



	Complete	Average	Ward
LC-3000	89.3	94.0	76.1
N-3000	98.9	99.9	96.2
HC-3000	98.1	99.5	94.3

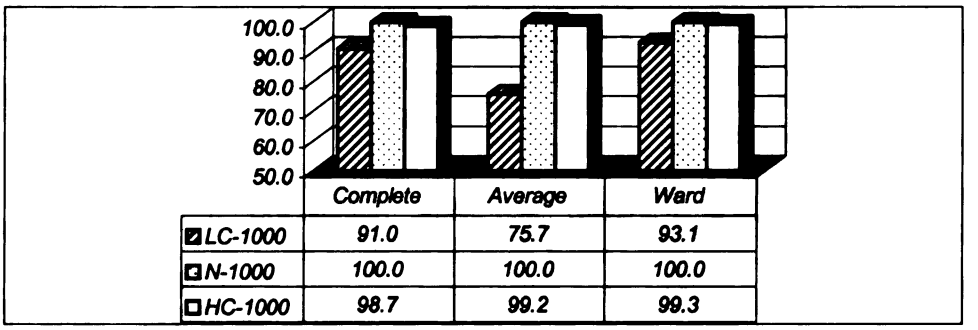
(d) 60 Items with sample size 3000

Figure 4.12. Ability and Guessing Combined Effect on Correct Classification for 3 Clusters of 3-Dimensional MS: Nonparametric Approach

decline in rates for the combination of lower ability and guessing were 16.1%, 17.9%, 13.4, and 12.0% for the 1000 sample size with 30 and 60 items, sample size 3000 with 30 and 60 items respectively.

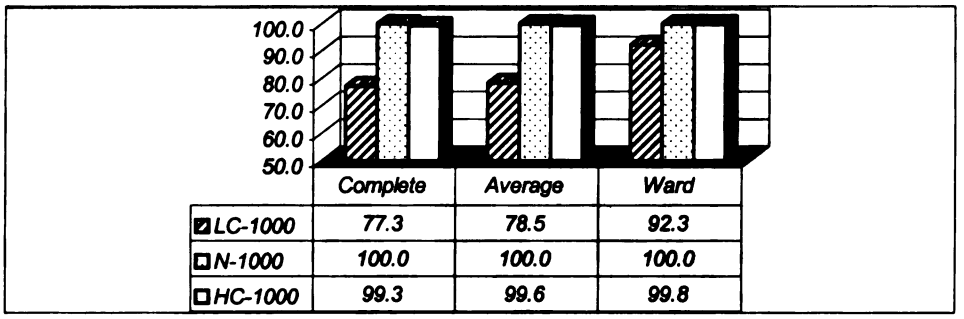
For the parametric approach, all HCA methods for classifying items into three dimensions seemed to be affected by the combination of ability and guessing factors. This was partly because all hit rates under the mean 0.0 condition were 100%. With the combination of higher ability and guessing, the hit rates dropped, but not to an unsatisfactory level. However, with the combination of lower ability and guessing, all perfect hit rates under the mean 0.0 condition became unsatisfactory, except those based on Ward's methods with sample size 1000. Figure 4.13 shows the effect of different abilities and guessing combinations on classification proportions for 3 dimensions with the parametric approach. The average declines for the combination of ability and guessing were 14% and 16% with 1000 and 3000 sample sizes respectively.

For the parametric approach, the summary of classification of 30 items into 6 clusters is excluded because all HCA methods with the mean 0.0 ability level were unsatisfactory (less than 87%). Figure 4.14 shows the effect of different abilities and guessing combinations on



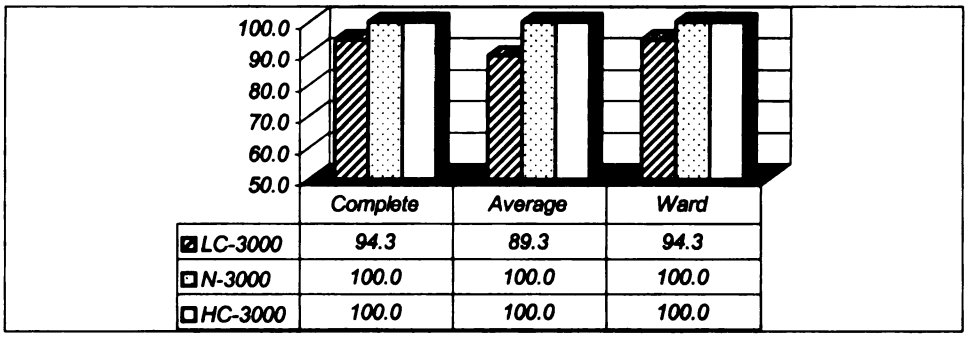
	Complete	Average	Ward
LC-1000	91.0	75.7	93.1
N-1000	100.0	100.0	100.0
HC-1000	98.7	99.2	99.3

(a) 30 items with sample size 1000



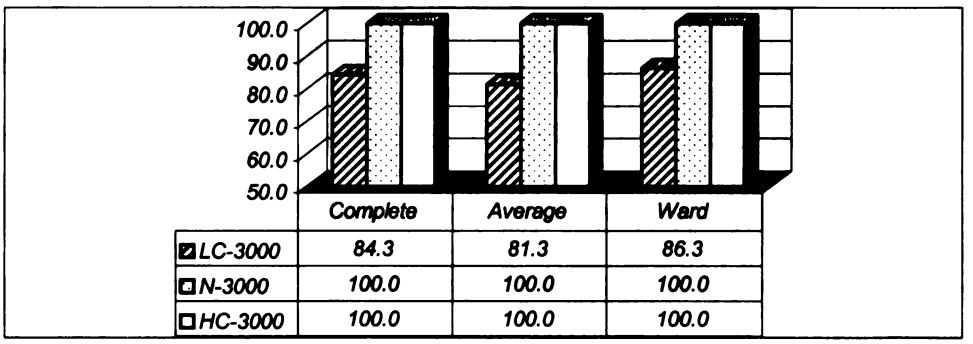
	Complete	Average	Ward
LC-1000	77.3	78.5	92.3
N-1000	100.0	100.0	100.0
HC-1000	99.3	99.6	99.8

(b) 60 Items with sample size 1000



	Complete	Average	Ward
LC-3000	94.3	89.3	94.3
N-3000	100.0	100.0	100.0
HC-3000	100.0	100.0	100.0

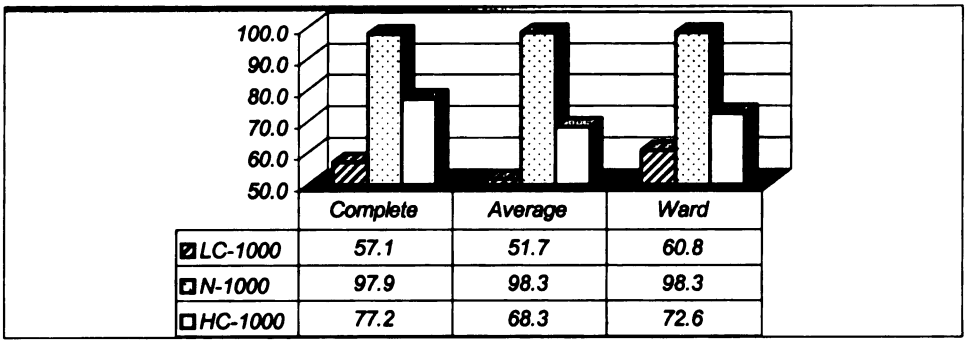
(c) 30 items with sample size 3000



	Complete	Average	Ward
LC-3000	84.3	81.3	86.3
N-3000	100.0	100.0	100.0
HC-3000	100.0	100.0	100.0

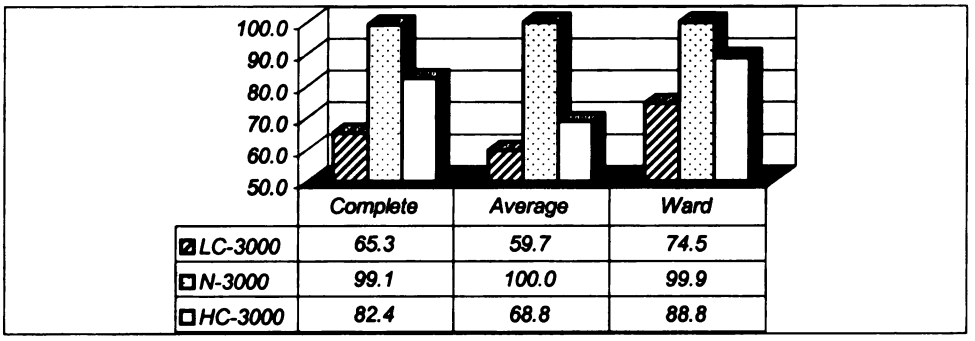
(d) 60 items with sample size 3000

Figure 4.13. Ability and Guessing Combined Effect on Correct Classification for 3 Clusters of 3-Dimensional MS: Parametric Approach



	Complete	Average	Ward
LC-1000	57.1	51.7	60.8
N-1000	97.9	98.3	98.3
HC-1000	77.2	68.3	72.6

(a) Sample size 1000



	Complete	Average	Ward
LC-3000	65.3	59.7	74.5
N-3000	99.1	100.0	99.9
HC-3000	82.4	68.8	88.8

(b) Sample size 3000

Figure 4.14. Ability and Guessing Combined Effect on Correct Classification for 6 Clusters of 3-Dimensional MS: Parametric Approach (60 items)

the proper classification (into 6 clusters) with 60 items for the parametric similarity measure. All HCA methods seemed to be affected by the different ability and guessing combinations. All satisfactory hit rates dropped to un-satisfactory. The average declines were 41.6% and 36.5% for the lower ability and guessing combination, and 25.5% and 19.7% for higher ability and guessing combined.

CHAPTER 5

SUMMARY AND DISCUSSION

Most IRT models currently in use are based on the assumption of unidimensionality. Since most achievement tests measure more than one skill, a test with items sampled from several content areas may not be sufficiently homogeneous to allow analysis using UIRT models. Even though many aspects of the MIRT model have been developed, there is not a single tool that can tell us how many dimensions a test has or how a test is structured in terms of dimensions or traits. This study evaluated the effectiveness of HCA/CCPROX developed by Roussos (1995) and the angular distance with HCA methods in identifying the number of dimensions and dimensionality structure. More specifically, a simulation study was conducted to determine the effectiveness of parametric and nonparametric proximity measures with HCA methods in correctly clustering items into their clusters. In addition, another set of simulations was conducted to determine the effect of guessing and ability levels on the efficiency of HCA methods for clustering items into their clusters.

5.1 Study I

5.1.1 Overall features of the two different approaches

Overall, the nonparametric approach was successful at clustering items correctly for the two- and three-dimensional APSS (i.e., classifying items into two clusters represented the two dimensions). The nonparametric approach also correctly clustered items for the orthogonal dimensions in the two- and three-dimensional MS case (i.e., classifying items into two clusters representing the two dimensions). However, the nonparametric approach was not successful at clustering items correctly into four clusters in the MS case (i.e., classifying items into four clusters represented as two approximately independent traits and two composite traits). The unsuccessful results could be due to the non-parametric approach's method of assessing dimensionality.

Figure 5.1 is an expanded version of Figure 3.3 (a). As indicated by Stout et al. (1996) and Zhang and Stout (1996), determination of whether an item pair belongs to the same dimension or not depends on the relative location (which is related to the conditional covariance) of each item in the two dimensional space. For example, items B and C belong to different clusters because they are located on opposite sides of the reference composite θ_{12} (the composite

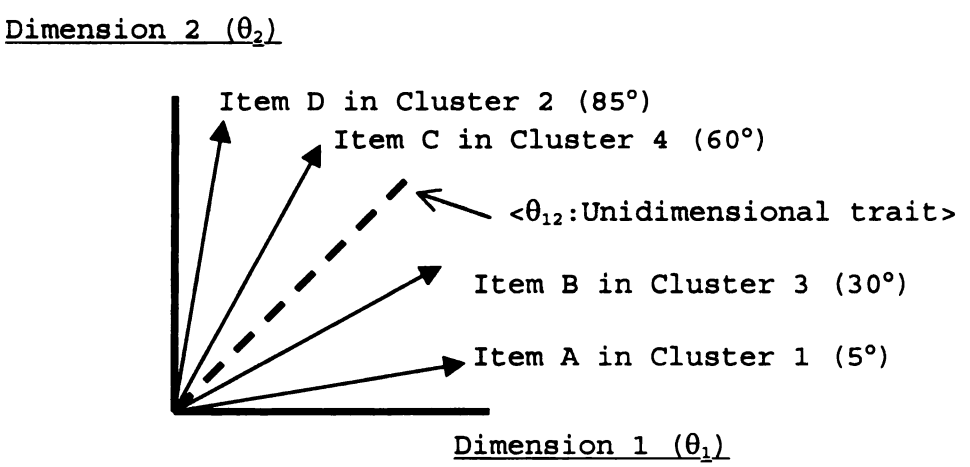


Figure 5.1. An Illustration of the Property of Non-parametric Approach to Cluster Items on Two-Dimensional MS

best measured by the total score). The nonparametric approach detects this well. On the other hand, even though item A and B belong to different clusters (A is an item of pure trait θ_1, and B is an item of the composite trait of θ_1 & θ_2), HCA with Pccor places the two items in the same cluster (dimension θ_1). This property of the nonparametric approach causes the poor hit rates for classifying correctly into four clusters. This implies, overall, that the nonparametric approach distinguishes orthogonal dimensions, but does not distinguish among approximately independent traits and composite traits.

One particular finding for the parametric approach was that all four HCA methods had perfect classification of two

and three clusters for the APSS with sample size 1000 and for the MS with both sample sizes. In fact, the parametric approach was successful in clustering items into their proper dimensions and their proper composite traits too. This implies that use of the angular distance as the proximity measure in the parametric approach clearly distinguished the difference not only between dimensions but also between clusters. For example, in (a) of Figure 3.3, the angle gap between each cluster was 10 degrees. The HCA methods with the angular distance as proximity measure were able to distinguish these 10-degree gaps in the two-dimensional mixed structure. In terms of dimensions (recovering two clusters), the angle gap between dimensions was also 10 degrees (40° to 50°). Thus whether the parametric approach worked or not seems to depend on the structure, more specifically on the angle gap between clusters.

It is clear that parametric and nonparametric approach have different properties for identifying a dimensional structure. That is, the nonparametric approach more likely identifies the clusters, which are representing the independent dimensions (traits) in a test, while the parametric approach more likely identifies the clusters wherever the angular distance defines membership of items

in a test. Some previous studies have compared clustering results for a standardized test based on the nonparametric and parametric approaches. However, these efforts did not take different properties of each methods into account. The procedures should have been compared in a different way. More specifically, they should have used the clustering results for identifying item memberships in terms of dimensions using nonparametric approach, and for identifying item membership in terms of combinations of traits (dimension and/or item content) using parametric approach. More investigations are needed using actual test data to find the uniqueness of the two approaches.

5.1.2. HCA Methods, Sample Size, and Number of Items in a Test and in a Cluster

Overall, the single link method was the worst in classifying proper items into clusters. The single link method has a tendency to induce chaining in the data (i.e., to form one single large cluster and another cluster of just a single case). This is the main reason why the proportions of correct classification for the single link method were just above the base rates. The average method seemed to be the best of all HCA methods for the non-parametric approach. Ward's method was the best one for the parametric

approach, but the difference in performance over other HCA methods was not large. Most HCA methods worked better with larger samples of examinees. One possible reason could be the stability of estimating parameters in MIRT with larger sample sizes, especially when sample size was larger than 2000. However, the difference between the 1000 and 3000 sample sizes were not large. HCA methods were satisfactory with both sample sizes. HCA methods tended to have higher hit rates with larger numbers of items in a test. More precisely, HCA methods tended to have higher hit rates with larger numbers of items in a dimension or a cluster. For the parametric approach, HCA methods showed the worst hit rates for 6 clusters with 3 dimensions under mixed structure tests (Table 9). These results reflect the instability of HCA methods for classifying relatively small numbers of items into relatively large numbers of clusters.

One particular finding was observed through simulation results. In general, the correct classification rates were higher when the numbers of items in a dimension are larger. However, the HCA methods with Pccor have tended to have lower hit rates for 60 items than for 40 items, for example, in the 2-D APSS model (Table 4.2). The complete link method with nonparametric approach clustered 95.8% and 85.8% correct classification for 40 and 60 items

respectively. This tendency was also observed in the previous study (Roussos et al., 1998). One possible hypothesis is that the conditional correlation may have a large standard error for the longer test because the frequencies in score categories is smaller than for the shorter test. The effect is especially strong when the sample size is small.

5.2 Study II

5.2.1 Effect of Guessing

Guessing slightly negatively affected the clustering results, especially when the sample size was small. In addition, guessing negatively influenced the hit rate when the number of clusters that HCA method had to classify was larger. For the parametric approach, the hit rates of 4 and 6 clusters for 2 and 3 dimensional mixed structure models were more likely affected by guessing. Again, these results reflect the instability of HCA methods for classifying relatively small number of items into relatively large number of clusters.

5.2.2 Effect of Different Ability Levels

Including different abilities levels has a slightly negative effect on the clustering results, especially for

the smaller sample sizes. In addition, this negative influence was greater when the mean ability level was lower, especially for the parametric approach. Item discrimination is traditionally used as a general indicator of item quality. In UIRT, precision of measurement provides another measure of quality at different levels of ability (θ) along the θ scale. In MIRT, the discriminating power of an item generally indicates how quickly the transition takes place from low probability to high probability of a correct response. The direction of an item vector related to MDISC and MDL indicates the weighted composite of abilities best measured by the item. Items with similar direction cosines or angles in the latent space measure similar weighted composites of abilities represented by that space.

The applications for which the parametric approach is appropriate are those that require comparisons of abilities needed for a correct response. Low ability examinees have limited distributions of scores in terms of the pattern of right answers. The pattern of direction cosines or angles of item vectors in a MIRT analysis for low ability distributions has somewhat failed to provide a means of identifying similarly and dissimilarly functioning sets of items in terms of what they measure.

5.2.3 Effect of Combination of Different Abilities and Guessing

Guessing with examinees of high ability had slightly negative affects on the clustering results, while guessing with low ability had significant effects. Low ability and guessing in combination had negative effects for the reason stated above. When the number of items was 20, the effects were the largest of all the conditions studied. As discussed before, this is because of the instability of HCA methods for relatively small numbers of items in a cluster.

5.3 Practical Implications

This study showed the relative effectiveness of two approaches for identifying the dimensionality structure of test under different conditions (different ability distributions, guessing, etc.). Since MDISC specifies a particular combination of abilities, the parametric approach is a more effective assessment tool for studying dimensionality than the nonparametric approach. Since parametric and non-parametric approaches work well with large sample sizes, these methods might work well for large-scale assessments like standardized achievement/attitude tests.

These two methods might help test developers and those who use tests to identify the structure underlying test content. Thus, the methods can be good tools for checking content validity. Suppose we have a nation wide standardized test and wish to analyze the dimensional structure. We can use DETECT to get an idea about how many dimensions are in the test. The results of DETECT can be confirmed with HCA/CCPROX (especially the average method) about which items belong to which orthogonal dimensions. With that particular number of dimensions identified using DETECT, we can run NOHARM to get discrimination parameter estimates and then compute the angular distances. Then we can analyze the structure of test contents by HCA (especially the Ward's method) with the parametric proximity measure.

If the test is relatively difficult, and relatively low examinees' ability level is expected, then some examinees are likely to guess when seeking the answers. In that case, caution is needed when using HCA with the parametric proximity measure because the simulation results have showed unstable classifications for those conditions. If the test is relatively difficult, and relatively high examinees' ability level is expected, then HCA with the parametric proximity measure may be useful to identify the structure of

contents. This use is supported by the stability of the correct classification rates with the higher ability level. The correct classification rates are stable, even if guessing is expected in that test. Thus, the use of HCA with the parametric proximity measure is highly recommended in that case as well.

The HCA with the angular distance proximities may be useful for checking test parallelism too. This method can give us empirical evidence to see if two parallel tests have intended parallel contents. Also, the mixed structure concepts developed in this study could be used in constructing multidimensional parallel test because the degree of multidimensional parallelism can be provided by the composite traits as discussed by Luecht and Miller (1992).

5.4 Limitations

This study has some limitations, as does any simulation study. Although the dimensional structures and other simulation factors were designed to be realistic, there are some gaps between simulations and real test data. For instance, the number of dimensions in a real test could be more than three. In addition, the number of clusters including composite traits could be greater than four or

six. The number of items in a cluster is not necessarily the same. Most perfect classifications obtained using some HCA methods with the parametric approach were partly due to clear differences between the simulated clusters. If the angles between clusters were less than 10 degrees, then the results might have been different. Another limitation is related to using the NOHARM program needed to estimate discrimination parameters to compute the angular distance. When specifying input options in NOHARM, the number of dimensions was assumed known already. In reality, nobody knows the real number of dimensions in a test. Thus, to apply the parametric approach in real settings, more caution is needed.

5.5 Future Research

This study investigated the comparative efficiency of HCA methods with parametric and nonparametric proximity measures in identifying the dimensional structure of a test. Although all the factors simulated and the dimensional structures designed were fairly realistic, there are more things to consider.

One thing that needs consideration is the effect of outliers on a cluster analysis. Outliers are always a problem in a statistical analysis. It is known that cluster

analyses are very sensitive to outliers. There are many ways to define outlier(s), but an outlier in cluster analysis implies a cluster with just one item (member). When this outlier is presented, the results of HCA methods in classifying proper items into clusters would be different from ones found in this study. Future work should examine the effect of outlier(s) on clustering results of HCA methods with parametric and nonparametric approaches. The existence of outlier(s) will likely have a negative influence on clustering results (Anderberg, 1973; Everitt, 1980; Milligan, 1980, 1981; Spath, 1980). Many validation studies (Milligan, 1980, 1981; Milligan, Soon & Sokol, 1983) confirmed that outliers confuse the decisions made about the number of clusters and cluster memberships in a cluster solution.

Milligan and Cooper (1985) examined the efficiency of thirty stopping rules in cluster analysis. They found that local maximum of the F statistic (Calinski & Harabasz, 1974) was the best way to identify number of clusters and members in each clusters. If the local maximum of the F statistic was applied to HCA methods (e.g., Ward's and the average method) with parametric and nonparametric proximity measures, it might be possible to correctly identify the appropriate number of dimensions.

In locating group members in proper groups, Multidimensional Scaling (MDS) Methods have been known to be very effective. Especially when the number of dimensions is two or three, MDS may be a good tool to identify membership in each dimension. Expanding simulation work on the effectiveness of parametric and nonparametric approaches using Multi-dimensional Scaling Methods for locating items into their proper dimension may lead to profitable procedures.

As discussed earlier, HCA methods with parametric proximity measures are very useful in validating the content of a test. These studies might include a comparison between classification from subject experts and HCA methods with the parametric approach.

Stout and his colleagues have tried to conduct dimensionality analysis with DETECT and HCA/CCPROX (nonparametric approach). However, these kinds of studies were limited to finding number of dimensions. The use of a combination of DETECT and HAC/CCPROX might not give us any idea of the structure of a test in terms of content (what the items try to measure). If dimensionality analysis on a standardized test with a combination of DETECT, nonparametric, and parametric approaches were conducted, the study might be

able to inform us of the number of dimensions, but also the content structure of the test.

References

Ackerman, T.A. (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.A. (1989). Unidimensional item response theory calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement, 13,* 113-127.

Aldenerfer, M.S. & Blashfield, R.K (1984). *Cluster analysis.* Beverly Hills, CA: Sage.

Aliaga, A. (1986). *Outliers in discriminant analysis.* Unpublished doctoral dissertation. University of Michigan.

Anderberg, M.R. (1973). *Cluster analysis for applications.* New York: Academic Press.

Anscombe, F.J. (1960). Rejection of outliers. *Technometrics, 2,* 123-148.

Barton, R.M. (1997). *Outliers & HACA: a study of outlier influence & a weighted MDA method for outlier.* Unpublished doctoral dissertation. University of Georgia.

Barnett, V., & Lewis, T. (1984). *Outliers in statistical data (3rd ed.).* New York: Wiley.

Bately, R., & Boss, M.W. (1993). The effects on parameter estimation of correlated dimensions and a distribution-restricted trait in a multidimensional item response model. *Applied Psychological Measurement, 17, 131-141.*

Blashfield, R.K. & Aldenerfer, M.S. (1978). The literature on cluster analysis. *Multivariate Behavioral Research, 13,* 271-295.

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

Carlson, J.E. (1987). *Multidimensional item response theory estimation: A computer program* (ACT Research Rep. No. 87-190. Iowa City IA: American College Testing Program.

Cattell, R.B. (1978). *The scientific use of factor analysis in behavioral and life sciences.* New York: Plenum.

Chen, W.H. & Thissen, D. (1994). *Local independence indices for item pairs using item response theory.* Unpublished manuscript.

Choi, S. (1996). *A response dichotomization technique for item parameter estimation of the multidimensional graded response model.* Unpublished doctoral dissertation. The University of Texas at Austin.

Cooper, M.C., and Milligan, G.W. (1985). *An empirical examination of the impact of outliers on four clustering methods.* Unpublished manuscript.

Cronbach, L.J. & Gleser, G.C. (1953). Assessing similarity between profiles. *The Psychological Bulletin, 50,* 456-473.

Cronbach, L.J. (1980). Validity on parole: How can we go straight? In W.B. Schreder (Ed.), *Measuring achievement: progress over a decade (New directions for testing and measurement, No.5, pp. 99-106)*. San Francisco: Jossey-Bass.

Drasgow, F., & Lissak, R. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology, 68*, 363-373.

Everitt, B.S. (1980). *Cluster analysis*. New York: John Wiley.

Florek, K., Lukaszewicz, J., Perkal, J., Steinhaus, H., & Zubrzycki. S. (1951). Sur La Liason et la Division des Points d'un Esemble Fini. *Colloquium Mathematics, 2*, 282-285

Fraser, C. (1988). *NOHARM: A computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory*. New South Wales, Australia: Center for Behavioral Studies, The University of New England.

Gordon, A.D. (1981). Classification: methods for cluster analysis. *Biometrics, December*, 623-637.

Gower, J.C., & Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of Lcassification, 3*, 5-48.

Hambleton, R.K., & Rovinelli, R.J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement, 10, 287-302*.

Harrison, D.A. (1986). Robustness of IRT parameter estimation to violation of the unidimensionality assumption. *Journal of Educational Measurement, 9*,139-164.

Holland, P.W., & Rosenbaum, P.R. (1986). Conditional association and unidimensionality in monotone latent variable models. *Annals of Statistics, 14, 1523-1543.*

Horn, J.J. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30, 179-185.*

Jain, A.K. & Dubes, R.C. (1988). *Algorithms for clustering data*. Englewood Cliffs, NJ: prentice Hall.

Junker, B.W., & Stout, W.F. (1994). Robustness of ability estimation when multiple traits are present with one trait dominant. In D. Laveault, B.D. Zumbo, M.E. Gessaroli & M.W. Boss (Eds.), *Modern theories of measurement: problems and issues (pp. 31-61)*. Ottawa, Canada: Edumetrics Research Group, University of Ottawa.

Kaiser, H.F. (1970). A second generation little jiffy. *Psychometrika, 35, 401-415.*

Kaufman, L. & Rousseeuw, P.J. (1990). *Finding groups in data*. John Wiley & Sons, New York.

Kim, H. (1994). *New techniques for the dimensionality assessment of standardized test data*. Unpublished doctoral dissertation. University of Illinois at Urbana-Champaign.

Kok, F. (1988). Item bias and test dimensionality. In R. Langeheine & J. Rost (Eds.), *Latent trait and latent class models (pp. 263-275)*. New York, NY: Plenum.

Korpi, M. & Haertel, E. (1984). *Locating reading test items in multidimensional space: an alternative analysis of test structure*. Paper presented at the annual meeting of American Educational Research Association. New Orleans, LA.

Lance, G.N., & Williams, W.T. (1967). A general theory of classificatory sorting strategies, I: Hierarchical systems. *Computer Journal, 9,* 373-380.

Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Lorr, M. (1983). *Cluster analysis for social scientists.* Sac Francisco, Jossey-Bass Publishers.

Luecht, R.M. & Miller, T.R. (1992). Unidimensional calibrations and interpretations of composite traits for multidimensional tests. *Applied Psychological Measurement, 16, 279-293.*

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the retrospective study of disease. *Journal of the National Cancer Institute, 22, 719-748.*

Mazor, K.M. (1993). *An investigation of the effects of conditioning on two ability estimates in DIF analysis when the data are two-dimensional.* Unpublished doctoral dissertation. The University of Massachusetts.

McDonald, R.P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34, 100-117.*

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

McKinley, R.L. & Kingston, N.W. (1988). *Confirmatory analysis of test structure using multidimensional IRT.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

McKinley, R.L., & Reckase, M.D. (1983). MAXLOG: A computer program for the estimation of the parameters of a multidimensional logistic model. *Behavior Research Methods & Instrumentation, 15,* 389-390.

McQuitty, L.L. (1960). Hierarchical linkage analysis for the isolation of types. *Educational and Psychological Measurement, 20,* 55-67.

McQuitty, L.L. (1966). Similarity analysis by reciprocal pairs for discrete and continuous data. *Educational and Psychological Measurement, 26,* 825-831.

Miller, T.R., & Hirsh, T.M. (1992). Cluster analysis of angular data in applications of multidimensional item response theory. *Applied Measurement in Education, 5,* 193-212.

Milligan, G.W. (1980). An examination of the effect of 6 types of error perturbation on fifteen clustering algorithms. *Psychometrika, 45,* 325-342.

Milligan, G.W. (1981). A Monte-Carlo study of thirty internal criterion measures for CA. *Psychometrika, 46,* 187-199.

Milligan, G.W. (1987). *A study of Beta flexible clustering method.* Unpublished working paper, WPS 87-61, College of Business, The Ohio State University.

Milligan, G.W. & Cooper, M.C. (1985). An examination of procedures for detecting the number of clusters in a data set. *Psychometrika, 50,* 159-179.

Milligan, G.W., Soon, S.C., & Sokol, L.M. (1983). The effect of cluster size, dimensionality and the number of clusters on recovery of true cluster structure. *IEEE transactions on Pattern Analysis and Machine Intelligence, 5,* 40-47.

Mislevy, R.J. & Bock, R.D. (1982). *BILOG: Item analysis and test scoring with binary logistic models (computer program).* Mooresville, Indiana: Scientific Software.

Nandakumar, R. (1994). Assessing dimensionality of a set of items. Comparison of different approaches. *Journal of Educational Measurement, 18,* 41-68.

Nandakumar, R., & Stout, W.F. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics, 18,* 41-68.

Oltman, P.K., Stricker, L.J., & Barrows, T. (1988). *Native language, English proficiency, and the structure of the Test of English as a Foreign language.* TOEFL Research, No. 27, (RR-88-26). Princeton, NJ: Educational Testing Service.

Oltman, P.K., Stricker, L.J., & Barrows, T. (1990). Analyzing test structure by multidimensional scaling. *Journal of Applied Psychology, 75,* 21-27.

Olson, J.F., Scheuneman, J., & Grima, A. (1989). *Statistical approaches to the study of item difficulty.* Educational testing Service Research Report 89-21.

Reckase, M.D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9,* 401-412.

Reckase, M.D. (1994). What is the "correct" dimensionality for a set of item response data? D. Laveault, B.D. Zumbo, M.E. Gessaroli, & MW. Boss. (Eds.) *Modern theories of measurement: Problems and issues, 87-92.*

Reckase, M.D. (1997a). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21, 25-35.*

Reckase, M.D. (1997b). *Constructs assessed by portfolios: How do they differ from those assessed by other educational tests.* Paper presented at the annual meeting of American Educational Research Association. Chicago, IL.

Reckase, M.D. (1997c). A linear logistic multidimensional model for dichotomous item response data. W.J. van der Linden & R.K. Hambleton (Eds.), In *Handbook of multidimensional item response theory, pp. 271-286.* Springer-Verlag New York Inc.

Reckase, M.D., Ackerman, T.A. and Carlson, J.E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement, 25,* 193-203.

Reckase, M.D., & McKinley, R.L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15,* 361-372.

Romesburg, H.C. (1990). *Cluster analysis for researchers.* Malabar, FL: Robert E. Krieger Publishing Company.

Rosenbaum, P.R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika, 49, 425-435.*

Roussos, L.A. (1992). *Hierarchical agglomerative clustering computer program users manual*. Unpublished manuscript, University of Illinois at Urbana-Champaign.

Roussos, L.A. (1995). *A new dimensionality estimation tool for multiple-item tests and a new DIF analysis paradigm based on multidimensionality and construct validity*. Unpublished doctoral dissertation. University of Illinois at Urbana-Champaign.

Roussos, L.A., Stout, W.F., & Marden, J.I. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. *Journal of Educational Measurement, 35, 1-30.*

Roznowski, M., Humphreys, L.G., & Davey, T. (1994). A simplex fitting approach to dimensionality assessment of binary data matrices. *Educational and Psychological Measurement, 54,* 263-283.

Roznowski, M., Tucker, L.R. & Humphreys, L.G. (1991). Three approaches to determining the dimensionality of binary items. *Applied Psychological Measurement, 15, 109-127.*

Sadek, R.F. & Huberty, C.J.. (1992). *On outlier influence in classification analysis.* Paper presented at the annual meeting of the American Educational Research Association. San Francisco, CA.

Sireci, S.G., & Geisinger, K.F. (1992). Analyzing test content using cluster analysis and multidimensional scaling. *Applied Psychological Measurement, 16,* 17-31.

Sokal, R.R. & Michener, C.D.. (1958). A statistical methods for evaluating systematic relationships. *University of Kansas Bulletin, 38,* 1409-1438.

Soon, S.C. (1987). *On detection of extreme data points in cluster analysis.* Unpublished doctoral dissertation. Ohio State U.

Spath, M. (1980). *Cluster analysis: Algorithms for data reduction and classification of objects.* Chichester, England: Ellis Horwood.

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

Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika, 52,* 589-617.

Stout, W. F., Habing, B., Douglas, J., Kim, H.R., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 19,* 331-354.

Sympson, J.B. (1978). A model for testing with multidimensional items. In D.J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference (pp. 82-98).* Minneapolis: University of Minnesota. Department of Psychology.

Wang, M.M. (1986). *Fitting a unidimensional model to multidimensional item response data* (ONR Rep. 042286). Iowa City IA: University of Iowa.

Ward, J.H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58,* 236-244.

Wise, S.L. (1983). Comparison of order analysis and factor analysis in assessing the dimensionality of binary data. *Applied Psychological Measurement,* 7, 311-321.

Wise, S.L., & Tatsuoka, M.M. (1986). Assessing the dimensionality of dichotomous data using modified order analysis. *Educational and Psychological Measurement, 46*, 295-301.

Vichi, M. (1985). Cluster analysis and the graphical approach for identification for two types of multivariate outliers. *Metron, 43*, 165-188.

Yen, W.M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-145.

Zhang, J., & Stout, W.F. (1995). *Theoretical results concerning DETECT.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Zhang, J., & Stout, W.F. (1996). *Theoretical index of dimensionality and its estimation.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, NY.

Zeng, L. (1989). *Robustness of unidimensional latent trait models when applied to multidimensional data.* Unpublished doctoral dissertation, University of Georgia.