2-2003

Conditional Covariance Based Subtest Selection for DIMTEST

Amy G. Froelich
Iowa State University, amyf@iastate.edu

Brian Habing
University of South Carolina

Conditional Covariance Based Subtest Selection for DIMTEST

Abstract

DIMTEST is a nonparametric hypothesis testing procedure designed to test the assumptions of unidimensionality and local independence for item response theory models. Several previous Monte Carlo studies have found using linear factor analysis to select the assessment subtest for DIMTEST results in a moderate to severe loss of power when the exam lacks simple structure, the ability and difficulty parameter distributions differ greatly, or the underlying model is non-compensatory. A new method of selecting the assessment subtest for DIMTEST based on the conditional covariance dimensionality programs DETECT and HCA/CCPROX is presented. Simulation studies show DIMTEST with this new selection method often has much higher power to detect multidimensionality than using linear factor analysis for subtest selection.

Keywords

DIMTEST, item response theory, unidimensionality, local independence, conditional covariance, linear factor analysis, HCA/CCPROX, DETECT

Disciplines

Statistics and Probability

Comments

This preprint was published as Amy Froelich and Brian Habing, "Conditional Covariance-Based Subtest Selection for DIMTEST", *Applied Psychological Measurement* (2008): 138-155, doi: 10.1177/0146621607300421

Conditional Covariance Based Subtest Selection for DIMTEST

Amy G. Froelich
Department of Statistics
Iowa State University

Brian Habing
Department of Statistics
University of South Carolina

Please send all correspondence to Amy G. Froelich, Department of Statistics, Iowa State University, 324 Snedecor Hall, Ames, IA 50011-1210. E-mail: amyf@iastate.edu

Conditional Covariance Based Subtest Selection for DIMTEST.

DIMTEST is a nonparametric hypothesis testing procedure designed to test the assumptions of unidimensionality and local independence for item response theory models. Several previous Monte Carlo studies have found using linear factor analysis to select the assessment subtest for DIMTEST results in a moderate to severe loss of power when the exam lacks simple structure, the ability and difficulty parameter distributions differ greatly, or the underlying model is non-compensatory. A new method of selecting the assessment subtest for DIMTEST based on the conditional covariance dimensionality programs DETECT and HCA/CCPROX is presented. Simulation studies show DIMTEST with this new selection method often has much higher power to detect multidimensionality than using linear factor analysis for subtest selection.

Index terms: DIMTEST, item response theory, unidimensionality, local independence, conditional covariance, linear factor analysis, HCA/CCPROX, DETECT.

Conditional Covariance Based Subtest Selection for DIMTEST.

Many of the most commonly employed statistical procedures for analyzing large scale standardized tests are based on the underlying model assumptions of unidimensionality (d=1) and local independence (LI). These include procedures for model fitting (e.g. BILOG and LOGIST), detection of differential item functioning (e.g. Mantel-Haenszel and SIBTEST), and equating. At an even more fundamental level, the assumption of unidimensionality is intimately related to whether or not it even makes sense to report a single score for an entire exam. The DIMTEST procedure (Stout, 1987; Nandakumar & Stout, 1993; Stout, Froelich, & Gao, 2001; Froelich & Stout, 2003) is one widely studied method that has been proposed for testing the hypothesis of d=1 and LI (e.g. Hattie, Krakowski, Rogers, and Swaminathan 1996; Seraphine, 2000).

The basic idea behind the DIMTEST procedure is to divide the items on the exam into a partitioning subtest (PT) and an assessment subtest (AT). If the exam is multidimensional these two subtests are ideally chosen to measure different composite abilities. The DIMTEST statistic calculated for these two subtests will then be large, causing the null hypothesis of d=1 and LI to be rejected. If the exam is unidimensional then the two subtests will of necessity measure the same ability. The DIMTEST statistic will be small and the null hypothesis of d=1 and LI will usually not be rejected.

Due to its simplicity of implementation, Stout (1987) suggested using principal axis factor analysis of the tetrachoric correlations (FAC) to determine appropriate AT and PT subtests if expert content analysis is not feasible. However, Hattie, et.al. (1996) and Seraphine (2000) (among others) have found that FAC can perform poorly in this capacity. In this paper we propose a method of item partitioning based on the same conditional covariance theory (Zhang and Stout, 1999a; Stout, Habing, Douglas, Kim, Roussos, and Zhang, 1996) as DIMTEST. In particular we combine the use of the conditional covariance based cluster analysis method HCA/CCPROX (Roussos, Stout, and Marden, 1998) with the DETECT statistic (Kim, 1994; Zhang and Stout, 1999b). The method is an easily automated expansion of the DIMTEST-HCA/CCPROX exploratory procedure discussed by Stout, Habing, et.al. (1996).

We begin by giving a brief overview of the theory of the conditional covariance based methods used in the paper: DIMTEST, HCA/CCPROX, and DETECT, with references to further details on each. The proposed method of determining the AT and PT clusters is then described. A simulation study then demonstrates that the new method significantly outperforms DIMTEST with FAC (DT-FAC) in a variety of more standard cases. Finally, the performance of the new method in the more complicated settings of Hattie, et.al. (1996) and Seraphine (2000) is briefly examined.

Conditional Covariance Based Dimensionality Assessment

Three assumptions that underly many of the most common IRT based procedures are local independence (LI), monotonicity (M), and unidimensionality (d=1). The commonly used Rasch and 3PL models both satisfy these assumptions, and they are needed for many standard procedures for equating and differential item functioning detection. Of course, it is not necessary for these three assumptions to hold in practice and a variety of models are available where one or more of the assumptions are relaxed (e.g. van der Linden & Hambleton, 1997). In the case of educational assessments it is often desired to assume monotonicity (more ability implies a larger chance of correct response) and local independence (the idea of a complete latent trait from Lord and Novick, 1968). The assumption of unidimensionality will often be unreasonable however. Consider a reading exam consisting of four paragraphs on different subject areas (Ancient History, US History, Physics, and Biology), each with six questions. In this case the exam would presumably have (nuisance) dimensions due to content area as well as the desired reading ability dimension.

A variety of models have been proposed for multidimensional exams, many of which belong to the family of LI, M, $d > 1$ generalized compensatory models (Zhang & Stout, 1999a). This family includes the compensatory logistic MIRT model (e.g. Reckase, 1997) of the form:

$$P[U_i = 1|(\theta_1, \dots \theta_d)] = c_i + \frac{1 - c_i}{1 + exp(-1.7(a_{i1}\theta_1 + \cdots + a_{id}\theta_d - \|\vec{a_i}\| b_i))} \qquad (1)$$

where U_i represents the response to item i, $\vec{\theta} = (\theta_1, \dots \theta_d)$ is the ability vector, $\vec{a_i}$ is the item discrimination vector, b_i is the item difficulty parameter, and c_i is the pseudo-guessing parameter. This class of models also includes the normal ogive based NOHARM model (McDonald, 1967).

Each item for an exam following a compensatory MIRT model can be represented geometrically (Reckase, 1997; Ackerman, 1996) by its discrimination vector $(a_{i1}, \dots, a_{id})$ as illustrated in Figure 1.

FIGURE 1 HERE

We say that the direction of an item's discrimination vector is the direction (or composite) best measured by the item. In Figure 1, Item 1 measures θ_1 alone. While Item 2 measures both θ_1 and θ_2, it best measures a composite ability that lies between the two main abilities. A set of items (such as the entire test, or the PT or AT subtests discussed below) can also be represented by a vector. Zhang and Stout (1999a) define the direction of best measurement of a test as the linear composite of the abilities that maximizes a multidimensional information function. The direction of best measurement of the test Θ_{TT} is a weighted average of the item discrimination vectors and can be thought of as giving an idea of what the test as a whole measures. While the unidimensional composite in the Zhang and Stout theory is usually not estimated directly by any procedure, we use Θ_{TT} to stand for its approximation by measures such as the total test score or the estimate from a unidimensional procedure such as BILOG (Habing & Roussos, in press).

The principle result from Zhang and Stout (1999a) is that we can recover information about the multidimensional structure represented by Figure 1 simply by finding the item pair conditional covariances based on the unidimensional Θ_{TT}:

$$CCOV_{i,l} = \int_{-\infty}^{\infty} Cov(U_i, U_l | \Theta_{TT} = \theta_{TT}) f(\theta_{TT}) d\theta_{TT}$$

If $CCOV_{i,l} > 0$ then items i and l are close in space and they measure similar ability composites; if $CCOV_{i,l} < 0$ then the direction of best measurement of items i and l are distant in space and they measure differing ability composites; and if $CCOV_{i,l} = 0$ then at least one of the items in the pair measures Θ_{TT}. In two dimensions this results in item pairs whose vectors lie on the same side of Θ_{TT} having a positive conditional covariance, and those that are on opposing sides having a negative conditional covariance. Further, the magnitudes of the $CCOV_{i,l}$ are related to the closeness of the items' directions of best measurement (along with their closeness to Θ_{TT} and the length of the discrimination vector).

Based on this theory, estimates of $CCOV_{i,l}$ gained from unidimensional methods can be used to analyze the underlying multidimensional structure of an exam. These ideas are implemented in the DIMTEST, HCA/CCPROX, and DETECT procedures. The heuristics of each of these procedures are described below so that the partitioning method described in this paper can be understood. A unified overview of these procedures can be found in Stout, Habing, Douglas, Kim, Roussos, and Zhang, (1996). More detailed references for the theory behind each of the individual methods are given below.

DIMTEST

The DIMTEST procedure (Stout, 1987; Nandakumar & Stout, 1993; Stout, Froelich, & Gao, 2001; Froelich & Stout, 2003) is designed to test the hypothesis that an exam satisfies the assumptions of $d = 1$ and LI. The procedure consists of two stages. First, the exam must be partitioned into two sets of items called the assessment subtest AT and the partitioning subtest PT. Ideally AT and PT will be chosen so that the AT items are dimensionally homogeneous and distinct from the PT items (see Figure 2.)

FIGURE 2 HERE

Once AT and PT have been chosen the second stage is to calculate an estimate of:

$$T^* = \sum_{i<l \in AT} \int_{-\infty}^{\infty} Cov(U_i, U_l | \Theta_{PT} = \theta_{PT}) f(\theta_{PT}) d\theta_{PT} \tag{2}$$

The positive bias in the estimate of T^* is then removed using a bootstrap technique and is compared to a reference normal distribution after standardization (Froelich & Stout, 2003). Notice that if AT and PT can be chosen as in Figure 2, Zhang and Stout's theory says that each of the conditional covariances being summed will be positive, resulting in a positive T^*. On the other hand, if the exam is unidimensional, AT and PT will necessarily coincide so that all of the items lie on θ_{PT}, making each of the summed covariances in T^* zero. Unfortunately a small value for T^* can also occur in a multidimensional exam if AT and PT are chosen poorly (as in Figure 3.)

FIGURE 3 HERE

The selection of AT and PT is thus vital to the performance of the DIMTEST procedure. Stout (1987) and most of the studies since then have used principal components factor analysis of the tetrachoric correlations on only a portion of the examinee pool to choose AT (a method abbreviated FAC). The DIMTEST statistic and hypothesis test was then calculated using the remainder of the examinees. It should be noted that earlier versions of DIMTEST used a portion of the PT subtest to correct for bias in the estimate of T^* (Stout, 1987, Nandakumar & Stout, 1993).

The bootstrap method used by Froelich & Stout (2003) is more powerful for detecting violations of d=1 and LI and is the method used here.

HCA/CCPROX

One method for determining clusters of items is HCA/CCPROX (Roussos, Stout, and Marden, 1998). This is an agglomerative hierarchical cluster analysis method. That is, each item starts as a separate cluster and at each stage the two clusters with the smallest distance between them are joined together, until at the final stage all of the items are massed together in a single cluster. The proximity measure $\rho_{ccov}(U_i, U_l)$ for a pair of items i and l is an estimate of:

$$-1 \times CCOV_{i,l} + constant \tag{3}$$

where the constant is added so that resulting proximity measure is non-negative.

Any two AT items or any two PT items in Figure 2 are on the same side of Θ_{TT} resulting in a positive conditional covariance, and after the sign reversal a small p_{ccov} value. An AT and PT item would fall on opposite sides of Θ_{TT}, resulting in a negative conditional covariance, and after the sign reversal a large ρ_{ccov} value. Thus items whose vectors are near each other will have small proximity values and those far away will have large proximity values. The distance between clusters used in HCA/CCPROX is the unweighted pair-group method of averages where the distance between two clusters is simply the average of the distances between all of the item pairs where one item is selected from each of the two clusters.

For the exam in Figure 2 the first stage of the cluster analysis would consist of all of the items in separate (singleton) clusters. At each stage it would join the AT and PT items together into progressively larger clusters until at the second to last stage one cluster would contain all of the AT items and the other would contain all of the PT items. At the final stage these two clusters would join together giving a list of all of the items on the exam. HCA/CCPROX virtually always recovers the structure of exams when they have approximate simple structure (all of the items measure only a few separate constructs). It does not however have any mechanism for telling you at what stage of the clustering you should stop. That is, since you cannot see Figure 2 you do not know that the two cluster solution is the correct one (as opposed to say the 1, 3, or 4 cluster solution). Also if the item directions of best measurement formed a fan shape instead of discrete clusters, HCA/CCPROX will correctly group those items at the extreme ends of the fan but have difficulty in assigning items in the middle of the fan. This is not unexpected, however, as the exam is clearly not able to be well represented by any set of discrete clusters.

DETECT

The DETECT procedure (Kim, 1994; Zhang and Stout, 1999b) is designed to find the optimal partitioning of the items on an exam under the assumption of approximate simple structure. Even when the exam does not have approximate simple structure the number of clusters found in the final partition should give a feeling for the dimensionality of the exam. The basic idea underlying DETECT is that when the items of an exam with simple structure are partitioned correctly then all of the items within a cluster will have a positive conditional covariance with each other, and a negative conditional covariance with items from any other cluster. The optimal partitioning $\mathcal{P}$ will thus maximize the estimate of:

$$D(\mathcal{P}, \Theta_{TT}) = \frac{2}{n(n-1)} \sum_{1 \leq i \leq ln} \int \delta_{i,l} CCOV_{i,l} \tag{4}$$

where $\delta_{i,l} = 1$ if items i and l are in the same cluster according to partition $\mathcal{P}$ and is -1 if they are in different clusters. The estimated quantity is called the DETECT statistic. Ideally the DETECT statistic would be applied to all possible partitionings of the items to find the partitioning of the items with the maximum value. However, in practice a genetic algorithm is applied to the initial item clusterings provided by HCA/CCPROX.

Applying the DETECT statistic to the partitioning in Figure 2 would result in a large value of $D(\mathcal{P}, \Theta_{TT})$. This is because all of the terms in the sum would be positive (the positive covariances between the AT items being multiplied by +1, the positive covariances between the PT items being multiplied by +1, and the negative covariances for pairs split between AT and PT being multiplied by -1). The DETECT statistic calculated for the partitioning in Figure 3 would be much smaller.

Unlike HCA/CCPROX, DETECT returns one fixed set of clusters so there is no question of how many clusters is appropriate. It does not provide any extra information that a cluster analysis dendogram would (such as how far the clusters are apart from each other and if there are any sub-clusters). Unlike DIMTEST, DETECT provides no test of hypotheses to determine if the cluster are indeed distinct from one another.

AT and PT Subtest Selection

As noted above, the first stage in the DIMTEST procedure is the partitioning of the exam into AT and PT subtests. Stout (1987) suggested two methods for partitioning the items into AT and PT subtests: (a) using the opinion of an expert in the content area, or (b) an automated partitioning method using standard principal axis factor analysis of the tetrachoric correlations. Hattie, et.al. (1996) found that DIMTEST with FAC (DT-FAC) did not perform well when the underlying model was non-compensatory and Seraphine (2000) found that DT-FAC did not perform well when the examinee ability distribution was different than that of the item difficulty parameters.

Evidence that FAC can be greatly improved upon was found by Froelich & Stout (2003) and Seraphine (2000). Both found that DIMTEST with expert opinion (DT-EX) greatly outperformed DT-FAC in some of the more complicated simulation settings (non-matching ability and difficulty parameters and lack of simple structure).

That DT-FAC can be improved upon should not be a surprise however. Linear factor analysis of tetrachoric correlations has long been known to perform poorly as a tool for dimensionality assessment of standardized test data (see Hattie (1985) for some early references).

As Stout (1987) noted FAC is "merely a data analytic technique for obtaining items that are as unidimensional as possible... the user is certainly free to substitute any other data analytic technique believed to effectively produce a unidimensional set of items." Hattie, et.al. (1996) studied the use of DT-FAC in the case where items followed a partially compensatory model. They compared the effectiveness of using FAC with a refined tetrachoric method and with a non-linear factor analysis (NOHARM). Neither of those other factor analytic methods performed appreciably better.

An alternative to using factor analytic methods would be to use methods based on the same conditional covariance based theory as DIMTEST. One such example is the DIMTEST-HCA/CCPROX Exploratory analysis described by Stout, Habing, et.al. (1996). This sequential combination of DIMTEST and HCA/CCPROX was used to determine which of the clusters found by HCA/CCPROX seemed dimensionally distinct from the other clusters, thus giving some idea as to how many clusters should be in the final cluster solution. However, this end goal is somewhat different than that required for choosing appropriate AT and PT clusters for DIMTEST.

Following Stout, Habing, et.al. (1996) and Froelich & Stout (2003) the choice of AT and PT partition should be made so that:

1. The AT items are relatively dimensionally homogeneous. In the geometric representation this means that the angle containing the AT items should be relatively narrow.

2. Θ_{AT} and Θ_{PT} should be as distinct as possible. In the geometric representation the angles between Θ_{AT} and Θ_{PT} should be as large as possible.

3. AT should contain at least four items and PT should contain at least half of the items on the test.

Our proposed method uses the conditional covariance based theory of Zhang and Stout (1999a), as implemented by HCA/CCPROX and DETECT, to find a partitioning into AT and PT subtests that should come as close as possible to satisfying the three requirements above.

1. Run HCA/CCPROX (Roussos, Stout, & Marden, 1996) on the exam. Each cluster containing between four and half of the items on the exam is classified as a potential AT subtest. For each potential AT subtest, the corresponding potential PT subtest is simply the set of remaining items.

2. Calculate the DETECT statistic (Zhang & Stout, 1999b) for each potential AT/PT subtest pair identified in Step 1. The potential AT/PT pair with the greatest DETECT statistic value is selected as the AT and PT subtests for calculating the DIMTEST statistic.

Since HCA/CCPROX begins by grouping the most dimensionally similar items, the set of potential AT subtests from step 1 should contain many small, dimensionally homogeneous clusters. If the exam is multidimensional, these clusters will eventually reach a size where they contain all the other items that are very dimensionally similar so that the resulting cluster is relatively dimensionally distinct from the remaining items on the test. The set of AT/PT partitions found in step 1 should thus include at least one which corresponds to the three guidelines given above. The DETECT statistic calculated in step 2 chooses the AT/PT partition from among these candidates. In particular, the DETECT statistic is made larger when the conditional covariances between items in AT are most positive, and when those between AT and PT items are most negative. This two stage procedure with HCA/CCPROX and DETECT

(HCD) therefore should result in the selection of a homogeneous AT subtest that is distinct from the remaining PT items.

Monte Carlo Simulation Study

To test the use of this new AT/PT selection method, a Monte Carlo simulation study was conducted. The simulation study is broken into two parts. The first study using the settings found in Stout (1987) and Froelich & Stout (2003). This is to compare the performance of DIMTEST with the new HCA/CCPROX-DETECT AT/PT selection method (DT-HCD) to DIMTEST with linear factor analysis (DT-FAC) under conditions commonly seen in simulation studies. The second study compares the performance of DT-HCD to DT-FAC under more nonstandard conditions with noncompensatory models (e.g. Hattie, et.al., 1996) and where the item difficulty and examinee ability distributions differ greatly (e.g. Seraphine, 2000). It is important to realize that by "standard conditions" that we do not mean that these are the conditions that will commonly occur in practice. We mean only that these conditions are similar to those most commonly found in simulation studies in the literature and are likely those under which most dimensionality assessment procedures should perform optimally.

Performance of DT-HCD Under Standard Conditions

In the first study, the settings used were chosen to replicate as much as possible the previous simulation studies conducted using FAC from Stout (1987) and Froelich & Stout (2003). DIMTEST with FAC (DT-FAC) was tested against DIMTEST with HCA/CCPROX-DETECT (DT-HCD) to determine both the Type I error rate (assuming d=1) and power (assuming $d > 1$) of the procedure with the different AT/PT selection methods.

For the Type I error study, item parameters from three real unidimensional tests, an Armed Services Vocational Aptitude Battery (ASVAB) Auto Shop test with 25 items (from Mislevy & Bock, 1984), an ACT Math (ACTM) test with 40 items (from Drasgow, 1987) and a SAT Verbal (SATV) test with 80 items (from Lord, 1968) were used. Table 1 gives the mean and standard deviation values for each item parameter and each test.

TABLE 1 HERE

Using these item parameters, examinee response data was simulated using the three parameter logistic (3PL) model (Birnbaum, 1968)

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + \exp[-1.7a_i(\theta - b_i)]}, \tag{5}$$

with examinee abilities θ generated from the standard normal distribution.

Three different levels of sample size (750, 1500, 2000 examinees) were generated using the 3PL model. The first (250, 500, 750) examinees respectively were used to select an AT and PT subtest (using either FAC or the HCA/CCPROX-DETECT method). The remaining examinees were used to calculate the DIMTEST statistic and to complete the hypothesis test.

All settings in the study (3 tests, 3 examinee levels, 2 AT/PT selection methods) were completely crossed giving 18 different conditions. For each condition, DIMTEST was run on 100 complete data sets and the number of rejections of the null hypothesis of d=1, LI recorded in Table 2. The nominal rate of rejection is $\alpha = 0.05$.

TABLE 2 HERE

For the power study, data was simulated using the two-dimensional compensatory logistic MIRT model given by the equation

$$P[U_i = 1|(\Theta_1 = \theta_1, \Theta_2 = \theta_2)] = c_i + \frac{1 - c_i}{1 + exp(-1.7(a_{i1}\theta_1 + a_{i2}\theta_2 - \|\vec{a_i}\|b_i))} \tag{6}$$

The examinee ability vector $\vec{\theta} = (\theta_1, \theta_2)$ in the model was generated from a bivariate normal distribution with means set to zero, variances set to one, and the correlation set to either 0.3, 0.7 or 0.9.

The item parameters for the multidimensional model were based on the item parameters from the three unidimensional tests (ASVAB, ACTM, and SATV) used in the Type I error study. The item difficulty and pseudo-guessing parameters b_i and c_i were taken directly from these three tests. The value of the item discrimination vector $\vec{a_i} = (a_{i1}, a_{i2})$ was based on the angle between the item's direction of best measurement and the θ_1 axis (denoted as β_i) and the value of the item discrimination parameter (denoted as a_i) from the ASVAB, ACTM and SATV tests. The vector was determined by the following equations

$$a_{i1} = a_i \cos(\beta_i) \quad \text{and} \quad a_{i2} = a_i \sin(\beta_i) \tag{7}$$

The angle between the item's direction of best measurement and the θ_1 axis was chosen according to three different multidimensional models. For the simple structure model, 2/3 of the items were randomly chosen to have angles of 0 degrees, while the remaining 1/3 of the items were chosen to have angles of 90 degrees. Thus, 2/3 of the items measure θ_1 only while the remaining 1/3 measure θ_2 only. For the approximate simple structure model, 2/3 of the items were randomly chosen to have angles between 0 and 30 degrees while the remaining 1/3 were chosen to have angles between 60 and 90 degrees. Thus, 2/3 of these items measure mostly θ_1 while 1/3 of the items measure mostly θ_2. For the no structure (fan) model, the angle for each item was randomly chosen to be between 0 and 90 degrees.

As in the Type I error study, three different levels of sample size (750, 1500, 2000 examinees) were generated using this multidimensional model. The first (250, 500, 750) examinees respectively were used to select an AT and PT subtest (using either FAC or the HCA/CCPROX-DETECT method). The remaining examinees were used to calculate the DIMTEST statistic and complete the hypothesis test.

All settings in the study (3 tests, 3 examinee levels, 2 AT/PT selection methods, 3 multidimensional models, 3 correlation levels) were completely crossed giving 162 different conditions. For each condition, DIMTEST was run on 100 complete data sets and the number of rejections of the null hypothesis of d=1, LI recorded in Table 3 for the simple structure model, Table 4 for the approximate simple structure model and Table 5 for the no structure (fan) model. The nominal rate of rejection is $\alpha = 0.05$.

TABLE 3 HERE

TABLE 4 HERE

TABLE 5 HERE

Performance of DT-HCD Under Non-standard Conditions

The purpose of the second simulation study was to determine the difference in power rates of DIMTEST with FAC and DIMTEST with the new subtest selection method in some of the more non-standard simulation settings used in Hattie, et.al. (1996) and Seraphine (2000). Hattie, et.al. (1996) studied the effect of the different multidimensional models on the ability of DIMTEST to reject the null hypothesis of d=1, LI. Specifically, they looked at generating two-dimensional data using a non-compensatory model from Sympson (1978). In this model, a decrease in one ability can be offset only by large increases in the second ability. In mathematical terms, the model is given by the equation

$$P(U_i = 1|\Theta_1 = \theta_1, \Theta_2 = \theta_2) = c_i + (1 - c_i) \prod_{d=1}^{2} \frac{1}{1 + exp(-1.7(a_{id}(\theta_d - b_{id})))}. \qquad (8)$$

Hattie, et.al. (1996) found the power of DIMTEST with FAC was severely reduced when the multidimensional data was generated using this non-compensatory model.

To determine the power of DT-FAC and DT-HCD using this non-compensatory model, the examinee ability vector $\vec{\theta} = (\theta_1, \theta_2)$ was generated using the bivariate normal distribution with means set to zero, variances set to one, and correlation set to either 0.3, 0.7, or 0.9. The item parameters in the model were selected according to the simple structure model used in the previous Monte Carlo study. The particular values for a_{i1}, a_{i2}, b_i and c_i were taken from the ACTM test. The item difficulty parameters b_{i1} and b_{i2} were determined using the following equations

$$b_{i1} = b_i \cos(\beta_i) \quad \text{and} \quad b_{i2} = b_i \sin(\beta_i) \tag{9}$$

where the angle β_i was set to 0 degrees for two-thirds of the test items and 90 degrees for the remaining one-third of the test items.

Using these parameters, examinee response data was then generated using Equation 8 for 1500 examinees. The first 500 examinees were used to select the AT/PT subtest using either the FAC method or the new method based on HCA/CCPROX-DETECT.

All settings in the study (3 correlations and 2 AT/PT selection methods) were fully crossed, giving 6 different conditions. For each condition, DIMTEST was run on 100 complete data sets and the number of rejections of the null hypothesis of d=1, LI recorded in Table 6. The nominal rate of rejection is $\alpha = 0.05$.

TABLE 6 HERE

Seraphine (2000) studied the effect on the performance of DIMTEST when the latent ability and item difficulty distributions differed. In the unidimensional case, the rejection rate of DIMTEST was near the overall nominal rate of $\alpha = 0.05$. However, the power of the procedure to reject the null hypothesis of d=1, LI was severely limited in some settings for the d=2 case. These settings were replicated using both DT-FAC and the new method based on HCA/CCPROX-DETECT. 1500 examinee response were generated on 50 items from the two-dimensional MIRT model from Equation 6 with the pseudo-guessing parameter c_i set to zero. The item difficulty parameters b_i were generated from a normal distribution with $\mu = 0$ and $\sigma = 1$ and $-2.0 \leq b_i \leq 2.0$. The item discrimination vector $\vec{a_i} = (a_{i1}, a_{i2})$ was determined using Equation 7 where a_i was generated from a lognormal distribution with $\mu = 1.13$ and $\sigma = 0.6$ with $0.4 \leq a_i \leq 2.0$. The angle $\beta_i, i = 1, \ldots, 5$ was set to $(0, 90, 15, 45, 60)$ degrees respectively and was repeated for items 6 through 50. Therefore, 20% of the items measured only θ_1, 20% of the items measured only θ_2 and the other 60% of the items measured some combination of θ_1 and θ_2. The examinee abilities θ_1 and θ_2 were generated from the normal distribution with identical means chosen from $\mu = (1.25, 1.50)$, identical standard deviations chosen from $\sigma = (1, 0.9, 0.8, 0.7)$ and with a correlation of 0.3.

For each data set, the first 500 examinees were used to select the AT/PT subtests using either FAC or the HCA/CCPROX-DETECT method. The remaining examinees were then used to calculate the DIMTEST statistic and complete the hypothesis test.

All settings in the study (2 means, 4 standard deviations, and 2 AT/PT selection methods) were fully crossed, giving 16 different conditions. For each condition, DIMTEST was run on 100 complete data sets and the number of rejections of the null hypothesis of d=1, LI recorded in Table 7. The nominal rate of rejection is $\alpha = 0.05$.

TABLE 7 HERE

Discussion of Results

Using the more standard simulation settings, there does not appear to be a significant differences in the rejection rate of d=1, LI between the two AT/PT selection methods when the underlying model is unidimensional. Both methods have an average rejection rate over all trials of slightly less than the nominal rate of $\alpha = 0.05$ (0.0422 for DT-FAC and 0.0433 for DT-HCD). Within each unidimensional test, both methods produced similar average rejection rates as well.

However, there does appear to be significant differences in the rejection rates between the two AT/PT selection methods when the underlying model is multidimensional. Both methods perform well in the easier cases of low correlation and in the simple structure model with medium correlation. And both methods perform equally poorly in the difficult cases of high correlations for the approximate simple structure and no structure models. However, DT-HCD has a rejection rate significantly higher than DT-FAC in the moderately difficult cases of high correlation and simple structure and medium correlation and approximate simple structure or no structure. For example, for the simple structure model with high correlation, the average rejection rate across all trials is 20.67 for DT-FAC and 45.67 for DT-HCD. If the lowest examinee level is removed, the average rejection rates change to 26 for DT-FAC and 60 for DT-HCD.

Under the non-standard simulation settings of differing ability and item difficulty distributions (e.g. Seraphine, 2000), there does appear to be significant differences in the rejection rates between the two AT/PT selection methods when the underlying model is multidimensional. When the new method based on HCA/CCPROX-DETECT is used, the power rates of DIMTEST are still extremely high, almost 100% in all cases. However, for DIMTEST with FAC, the power rates decrease sharply, as was seen in the simulation results in Seraphine (2000).

Under the non-standard simulation setting of a non-compensatory multidimensional model (e.g. Hattie, et.al., 1996), there does not appear to be a large differences in the rejection rates between the two AT/PT selection methods, with DT-HCD performing somewhat worse. Both methods have high power rates when the correlation between dimensions is 0.3. However, neither method has power rates much different than the nominal rate of $\alpha = 0.05$ when the correlation is either 0.7 or 0.9.

Although the results show low power rates for DIMTEST when the items come from a non-compensatory model with high correlation between abilities, these findings are not surprising. The theory of conditional covariances from Zhang & Stout (1999a), on which DIMTEST and our AT/PT selection method are based, assumes a generalized compensatory model holds for the multidimensional data. Little is known about the behavior of these conditional covariances if another multidimensional model is assumed. In addition, the assumption of the non-compensatory model (a decrease in one ability can be offset only by large increases in the second ability) might not be valid if the correlation between the two abilities is large. Finally, little research has been done to equate item parameters for the compensatory and non-compensatory models. This leaves open the possibility that the item parameters used in our study are not realistic for non-compensatory models.

Conclusions

DIMTEST is a nonparametric hypothesis testing procedure designed to test the assumptions of d=1 and LI for item response theory models. Several previous Monte Carlo studies (e.g. Froelich & Stout, 2003; Seraphine, 2000; Hattie, et.al., 1996) have found using linear factor analysis to select the AT and PT subtests for DIMTEST results in a moderate to severe loss of power when the underlying model is non-compensatory or lacks simple structure, or when the ability and difficulty parameter distributions differ greatly. Using a new method of AT/PT subtest selection based on the conditional covariance programs HCA/CCPROX and DETECT, DIMTEST was shown to have much greater power when the underlying model lacked simple structure, or when the item difficulty and ability distributions were different. However, this new method did not seem to impact the power of DIMTEST when the items were generated using a non-compensatory multidimensional model. As noted in Hattie, et.al. (1996), DIMTEST appears to have low rejection rates in this setting, especially when the examinee abilities are highly correlated. These findings indicate the need for further research in this last case.

References

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

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, *Statistical theories of mental test scores*, 395-479. Menlo Park, CA: Addison-Wesley Publishing Company.

Drasgow, F. (1987). A study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72*, 19-30.

Froelich, A.G. & Stout W. (2003). A new bias correction method for the DIMTEST procedure. Unpublished manuscript submitted for publication.

Gessaroli, M.E., & De Champlian, A. (1996). Using an appropriate chi-square statistic to test the number of dimensions underlying the responses to a set of items. *Journal of Educational Measurement, 2*, 157-179.

Habing, B. & Roussos, L.A. (in press). On the need for negative local item dependence. Accepted to appear in *Psychometrika*.

Hattie, J., Krakowski, K., Rogers, J., & Swaminathan, H. (1996). An assessment of Stout's index of essential dimensionality. *Applied Psychological Measurement, 20*, 1-14.

Hattie, J.A. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*, 139-164.

Kim, H.R. (1994). New techniques for the dimensionality assessment of standardized test data. Unpublished Doctoral Dissertation. University of Illinois at Urbana-Champaign, Department of Statistics.

Lord, F.M. (1968). An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28*, 989-1020.

Lord, F.M., & Novick, M.R. (1968). *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley Publishing Company, Inc.

McDonald, R.P. (1967). *Nonlinear Factor Analysis.* Psychometric Monographs, No. 15.

Mislevy, R.J., & Bock, R.D. (1984). Item operating characteristics of the Armed Services Aptitude Battery (ASVAB), Form 8A, (Technical Report N00014-83-C-0283). Washington, DC: Office of Naval Research.

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

Reckase, M.D. (1997). A linear logistic model for dichotomous item response data. In W.J. van der Linden & R.K. Hambleton (Eds.), Handbook of Modern Item Response Theory, 271-286.

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

Seraphine, A.E. (2000). The performance of DIMTEST when latent trait and item difficulty distributions differ. *Applied Psychological Measurement, 42*, 82-94.

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

Stout, W., Froelich, A.G., & Gao, F. (2001). Using resampling to produce an improved DIMTEST procedure. In A. Boomsma, M.A.J. van Dujin, & T.A.B. Snijders, *Essays on Item Response Theory*, 357-375.

Stout, W., 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, Computerized Adaptive Testing Laboratory.

van der Linden, W.J., & Hambleton, R.K. (eds.) (1997). *Handbook of Modern Item Response Theory.* New York: Springer.

Zhang, J., & Stout, W. (1999a). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika, 64*, 129–152.

Zhang, J., & Stout, W. (1999b). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64*, 213–249.

Figure 1: Graphical Representation of Items

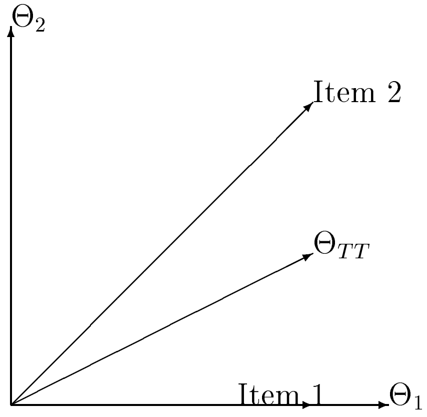

Figure 2: Good Choice for AT

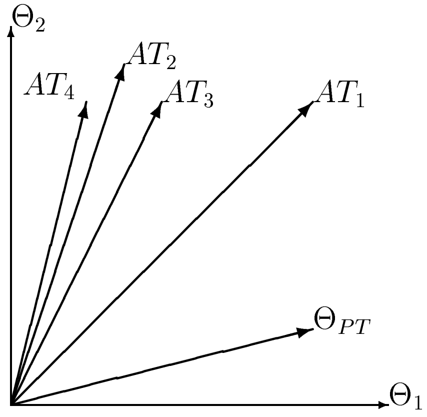

Figure 3: Poor Choice for AT

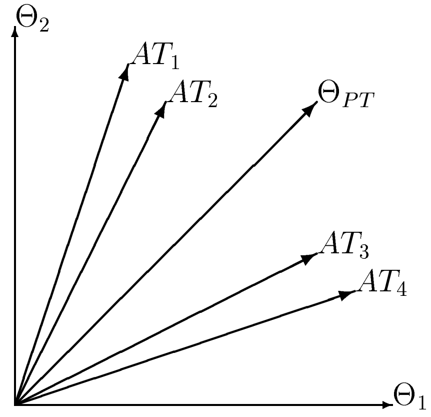

Table 1

Mean and Standard Deviation of Three Unidimensional Tests

Test	ASVAB			ACTM			SATV		
Parameter	a	b	c	a	b	c	a	b	c
Mean	1.22	0.09	0.20	1.09	0.50	0.14	1.07	0.58	0.16
St. Dev.	0.70	0.72	0.04	0.35	0.61	0.04	0.40	0.88	0.05

Table 2

Number of Rejections (d=1) for DIMTEST

Test	ASVAB			ACTM			SATV		
Examinees	750	1500	2000	750	1500	2000	750	1500	2000
DT-FAC	2	2	5	4	3	8	3	3	8
DT-HCD	5	3	4	9	1	7	3	2	5

Table 3

Number of Rejections (d=2) for DIMTEST - Simple Structure Model

	Test	ASVAB			ACTM			SATV		
Corr	Examinees	750	1500	2000	750	1500	2000	750	1500	2000
0.3	DT-FAC	100	100	100	100	100	100	100	100	100
	DT-HCD	100	100	100	100	100	100	100	100	100
0.7	DT-FAC	95	100	100	99	100	100	95	100	100
	DT-HCD	94	100	100	97	100	100	99	100	100
0.9	DT-FAC	4	20	46	15	22	29	11	13	26
	DT-HCD	22	43	56	15	54	72	14	55	80

Table 4

Number of Rejections (d=2) for DIMTEST - Approximate Simple Structure Model

	Test	ASVAB			ACTM			SATV		
Corr	Examinees	750	1500	2000	750	1500	2000	750	1500	2000
0.3	DT-FAC	99	100	100	98	100	100	100	100	100
	DT-HCD	100	100	100	100	100	100	100	100	100
0.7	DT-FAC	42	83	98	32	66	67	33	46	46
	DT-HCD	55	86	99	33	70	91	55	96	100
0.9	DT-FAC	5	5	9	4	3	7	3	8	15
	DT-HCD	4	12	16	7	11	11	2	14	18

Table 5

Number of Rejections (d=2) for DIMTEST - No Structure (Fan) Model

	Test	ASVAB			ACTM			SATV		
Corr	Examinees	750	1500	2000	750	1500	2000	750	1500	2000
0.3	DT-FAC	94	100	100	92	99	100	85	100	100
	DT-HCD	90	99	100	95	100	100	97	100	100
0.7	DT-FAC	17	46	49	14	31	28	3	5	8
	DT-HCD	22	63	73	10	62	86	15	55	83
0.9	DT-FAC	3	7	2	1	3	7	7	9	6
	DT-HCD	3	4	4	4	5	3	2	1	4

Table 6

Number of Rejections (d=2) for DIMTEST

Non-compensatory Model

Correlation	0.3	0.7	0.9
DT-FAC	95	19	5
DT-HCD	85	13	0

Table 7

Number of Rejections (d=2) for DIMTEST

Multidimensional Model

μ	σ	1.0	0.9	0.8	0.7
1.25	DT-FAC	99	80	52	32
	DT-HCD	100	100	100	100
1.50	DT-FAC	90	84	46	33
	DT-HCD	100	100	100	95