

Refinements of the DIMTEST Methodology for Testing Unidimensionality and Local Independence

Amy G Froelich  
Department of Statistics  
Iowa State University  
324 Snedecor Hall  
Ames, IA 50011-1210

Brian Habing  
Department of Statistics  
University of South Carolina  
Columbia, SC 29208

Copies of this paper can be obtained in pdf format on the first author's web site  
<http://www.public.iastate.edu/~amyf/papers.html>

## Refinements of the DIMTEST Methodology for Testing Unidimensionality and Local Independence

Many statistical procedures for analyzing large scale testing situations are based on the underlying model assumptions of unidimensionality and local independence. These include procedures for model fitting (such as BILOG and LOGIST) and procedures for the detection of DIF (such as Mantel-Haenszel and SIBTEST). The DIMTEST procedure (Stout, 1987) has been widely applied to the testing of this hypothesis and has been shown to be quite effective in a variety of circumstances.

Recently, a new bias correction method has been developed for the DIMTEST procedure (Froelich, 2001) which removes the need for the AT2 subtest of items. The bias correction method is based on a parametric bootstrap procedure using non-parametrically estimated item response functions. This new DIMTEST procedure has been shown to have Type I error rates near the nominal rate of  $\alpha = 0.05$  and increased power over the original procedure in almost all circumstances.

With this new bias correction method, the new DIMTEST procedure divides the test items into just two groups, a partitioning subtest (PT) and an assessment subtest (AT). The additional subtest (AT2) is no longer required for bias correction. DIMTEST then tests the hypothesis that the AT subtest measures the same construct as the PT subtest. DIMTEST is most powerful when the subtests are chosen so that the AT items are dimensionally homogeneous, and so that the constructs best measured by the AT and PT subtests are as distinct as possible. This new version of DIMTEST has more power to detect multidimensionality because it effectively has a larger number of items in the exam. The items that would have been assigned to the AT2 subtest are now assigned to either AT or PT instead.

The two subtests AT and PT are best chosen by expert opinion (based on a table of test specifications, for example). However, it would also be desirable to have an automated method of selecting the two subtests based on the examinee response data. In almost all of the past research studies concerning the DIMTEST procedure (Stout, 1987, Nandakumar & Stout, 1993, Hattie, Krakowski, Rogers & Swaminathan (1996), and Froelich, 2001) the AT and PT subtests have been selected using linear factor analysis based on the tetrachoric correlation matrix (called FAC in the DIMTEST program). Several researchers (see for example, Hulin, Drasgow, & Parsons, 1983; McDonald, 1981) have found linear factor analysis unsuitable for dimensionality assessment in a variety of circumstances. In particular, Froelich (2001) found, even for the new DIMTEST procedure, using linear factor analysis to select the AT and PT subtests produces diminished hypothesis testing power in some multidimensional models and little power in other multidimensional models. It is important to note here that the use of FAC is not a necessary part of the DIMTEST procedure. Any other reasonable method of selecting the AT subtest can always be employed. The simulation study in Froelich (2001) showed that the weaknesses in FAC do not demonstrate a lack in the DIMTEST statistic itself, but only demonstrate the need for a better method of mimicking “expert opinion.”

The basis for the DIMTEST procedure is the item pair conditional covariances. Two other procedures have been developed using these item pair conditional covariances, DETECT (Kim, 1994 and Zhang & Stout, 1999a) and HCA/CCPROX (Roussos, Stout, & Marden, 1998). The HCA/CCPROX procedure performs a hierarchical cluster analysis of the test items based on a proximity matrix calculated using the item pair conditional covariances. The DETECT procedure uses a genetic algorithm to determine the number of dimensions for a test and to divide the test items into distinct clusters. The optimal solution of the DETECT procedure is the number of dimensions and the particular item clusters that maximize the DETECT statistic. Both of these procedures are exploratory in nature; neither procedure has a hypothesis test associated with it. However, these procedures are a natural choice to select the AT subtest for the DIMTEST

procedure.

In this paper, we seek to address the concerns of Froelich (2001) and other DIMTEST studies by replacing the linear factor analysis method of selecting the AT and PT subtests with a method based on a combination of the DETECT and HCA/CCPROX procedures. Using this new method, the simulation study of Froelich (2001) is also updated and expanded.

## 1 New DIMTEST procedure

The following section explains the new version of the DIMTEST procedure. A complete discussion of the theory and steps in the new DIMTEST procedure can be found in Froelich (2001).

**Step 1.** Choose  $m$  items for the AT Subtest. (Since there is no AT2 Subtest, we simply denote the assessment subtest as AT). When multidimensionality is present in the test data, the goal is to choose AT so that the items are dimensionally homogeneous on a dimension distinct from the direction of best measurement of the remaining test items.

**Step 2.** Place the remaining  $(n - m)$  items in the PT Subtest. Define the  $k$ th examinee subgroup as all examinees whose total score on the PT Subtest, denoted as  $Z_{PT}$ , is equal to  $k$ . Let  $U_{ij}^{(k)}$  denote the response of the  $j$ th examinee from subgroup  $k$  to the  $i$ th assessment item and let  $J_k$  denote the number of examinees in subgroup  $k$ . An examinee subgroup  $k$  is eliminated from the DIMTEST statistic calculation if  $J_k$  is less than a specified minimum size (typically the minimum size of  $J_k$  ranges from 2 to 20). Denote the number of subgroups used in the calculation of the DIMTEST statistic as  $K$ . For each examinee subgroup  $k$ , calculate the following quantities:

$$Y_j^{(k)} = \sum_{i=1}^m U_{ij}^{(k)}, \quad \bar{Y}^{(k)} = \frac{1}{J_k} \sum_{j=1}^{J_k} Y_j^{(k)}, \quad \hat{p}_i^{(k)} = \frac{1}{J_k} \sum_{j=1}^{J_k} U_{ij}^{(k)},$$

$$\hat{\sigma}_k^2 = \frac{1}{J_k} \sum_{j=1}^{J_k} (Y_j^{(k)} - \bar{Y}^{(k)})^2, \quad \text{and} \quad (1)$$

$$\hat{\sigma}_{U,k}^2 = \sum_{i=1}^m \hat{p}_i^{(k)} (1 - \hat{p}_i^{(k)}). \quad (2)$$

**Step 3.** For each examinee subgroup  $k$ , calculate the statistic

$$T_{L,k} = \hat{\sigma}_k^2 - \hat{\sigma}_{U,k}^2 = 2 \sum_{i < l \in AT1} \widehat{Cov}(U_i, U_l | Z_{PT} = k) \quad (3)$$

where  $\widehat{Cov}(U_i, U_l | Z_{PT} = k)$  is the usual estimate of the covariance between two items conditioned on the set of all examinees with PT Subtest score  $k$  (Gao, 1997).

**Step 4.** The asymptotic variance of  $T_{L,k}$ , denoted as  $S_k^2$ , is calculated as

$$S_k^2 = \frac{(\hat{\mu}_{4,k} - \hat{\sigma}_k^4) - \hat{\delta}_{4,k}}{J_k}, \quad (4)$$

where

$$\hat{\mu}_{4,k} = \frac{1}{J_k} \sum_{j=1}^{J_k} (Y_j^{(k)} - \bar{Y}^{(k)})^4 \quad \text{and} \quad \hat{\delta}_{4,k} = \sum_{i=1}^m \hat{p}_i^{(k)} (1 - \hat{p}_i^{(k)}) (1 - 2\hat{p}_i^{(k)})^2.$$

The DIMTEST statistic  $T_L$  is given by

$$T_L = \frac{\sum_{k=1}^K T_{L,k}}{\sqrt{\sum_{k=1}^K S_k^2}} \quad (5)$$

**Step 5.** For each test item, calculate an estimate of the item's unidimensional IRF using a kernel smoothing procedure adapted from Ramsay(1991) and Douglas (1997). See Froelich (2001) for a complete description.

**Step 6.** Generate examinee responses to all test items using the estimated IRFs calculated in Step 5. See Froelich (2001) for a complete description.

**Step 7.** Using the generated data set from Step 6 and the same (AT,PT) partition as the original data, calculate another DIMTEST statistic according to Steps 2 through 4 and denote this statistic as  $T_G$ .

**Step 8.** To reduce the random variation in the  $T_G$  statistic, Steps 6 and 7 are repeated  $N$  times and the average of the  $N$   $T_G$  values, denoted as  $\bar{T}_G$ , is calculated. Under the assumption of unidimensionality, the final DIMTEST statistic,

$$T = \frac{T_L - \bar{T}_G}{\sqrt{1 + 1/N}} \quad (6)$$

has an asymptotically standard normal distribution under certain regularity conditions as the number of items and the number of examinees tends to infinity (see Froelich (2001) for a proof). Thus, the null hypothesis of unidimensionality is rejected at level  $\alpha$  if  $T$  is larger than the  $100(1 - \alpha)$ th percentile of the standard normal distribution.

## 2 Simulation Study from Froelich (2001)

The purpose of the simulation study from Froelich (2001) was to assess the performance of the new version of DIMTEST. The simulation study was split into two parts according to whether the data was unidimensional (measuring Type I error) or multidimensional (measuring power).

### 2.1 Type I Error Study

Examinee response data were simulated for the Type I error study using the unidimensional three parameter logistic (3PL) model

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + \exp[-1.7a_i(\theta - b_i)]}, \quad (7)$$

with examinee abilities  $\theta$  generated from the  $N(0, 1)$  distribution. The  $a$ ,  $b$  and  $c$  item parameters used in the study were estimated from three real tests: an Armed Services Vocational Aptitude Battery (ASVAB) Auto Shop test with 25 items (from Mislavy & 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).

For all DIMTEST runs, the program FAC, an exploratory linear factor analysis program based on tetrachoric correlations, was used to select the AT subtest (see Stout, 1987 and Nandakumar & Stout, 1993). For each trial, the size of the AT subtest was allowed to vary between four and one-half the number of test items.

Four levels of the number of examinees were used in the study (750, 1000, 1500, 2000). For each DIMTEST run, the first (250, 350, 500, 750) examinees were used to determine the AT Subtest and the remaining (500, 650, 1000, 1250) examinees were used to calculate the DIMTEST statistic respectively.

For all DIMTEST runs, the minimum cell size used in calculating the DIMTEST statistic was set to two and the number of resamplings of the data was set to  $N = 50$ .

All levels of the design: 3 tests and 4 examinee levels were fully crossed, giving 12 different unidimensional models. Each model was simulated 400 times and the number of rejections of the null hypothesis of unidimensionality recorded. The rate of rejection per 100 DIMTEST runs for each simulated model is given in Table 1. The nominal rate of rejection is  $\alpha = 0.05$ .

Table 1: DIMTEST: Type I Error Results

Test	ASVAB	ACTM	SATV
J = 750	1.50	2.25	2.00
J = 1000	3.00	4.50	1.75
J = 1500	4.00	4.75	2.75
J = 2000	2.00	3.00	3.50

## 2.2 Power Study

Examinee response data were simulated for the power study using the two dimensional version of the three parameter logistic IRT model

$$P_i(\boldsymbol{\theta}) = c_i + \frac{1 - c_i}{1 + \exp[-1.7\mathbf{a}_i^T(\boldsymbol{\theta} - \mathbf{b}_i)]}, \quad (8)$$

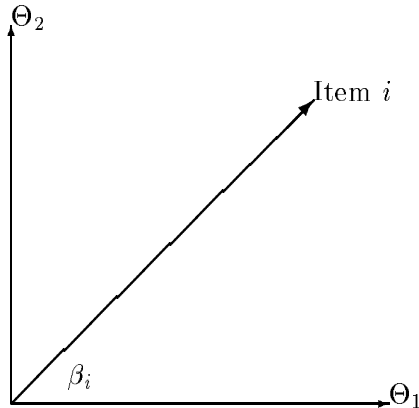
with examinee abilities generated from the bivariate normal distribution  $N(0, 0, 1, 1, \rho)$  where  $\rho = 0.3$  or  $0.7$ .

Three different two dimensional models were used for the IRFs: simple structure, approximate simple structure, and no structure. Let  $\beta_i$  denote the angle between the  $i$ th item's direction of best measurement and the  $\Theta_1$  axis as shown in Figure 1.

For the simple structure model, half of the test items were randomly chosen to have  $\beta_i = 0^\circ$ , while the other half of the items were assigned  $\beta_i = 90^\circ$ . For the approximate simple structure model, half of the test items were randomly chosen to have  $0^\circ \leq \beta_i \leq 20^\circ$ , while the other half of the items were assigned to have  $70^\circ \leq \beta_i \leq 90^\circ$ . The particular value of  $\beta_i$  for each item  $i$  was generated from the Uniform(0,20) or from the Uniform(70,90) distribution respectively. For the no structure model, the angle  $\beta_i$  for each item  $i$  was randomly generated from the Uniform(0,90) distribution.

The item parameters  $(a_{1,i}, a_{2,i}, b_{1,i}, b_{2,i}, c_i)$  for each item varied depending on the two dimensional model used for the IRFs. Let  $a_i$  denote the value of the item discrimination parameter for item  $i$  from the three tests used in the Type I error study: ASVAB, ACTM, and SATV. The item

Figure 1: Angle for Power Simulations



discrimination parameters  $a_{i,1}$  and  $a_{i,2}$  were determined for each item  $i$  by the relationships

$$a_{i,1} = a_i \cos(\beta_i) \quad \text{and} \quad a_{i,2} = a_i \sin(\beta_i)$$

For the simple structure model, the item difficulty parameters  $b_1$  and  $b_2$  and the guessing parameters  $c$  were taken from the three unidimensional tests used in the Type I error study. For the other two multidimensional models, the item difficulty parameters  $b_1$  and  $b_2$  were generated from the standard normal distribution and truncated at  $-1.5$  and  $1.5$ , while the  $c$  parameters were taken from the three unidimensional tests used in the Type I error study.

For all DIMTEST runs, the program FAC was used to select the AT subtest. Again, the number of items in AT was allowed to vary between 4 and one-half of the test items for each trial. To determine the power of the DIMTEST procedure when the AT Subtest is correctly specified for each trial, the DIMTEST runs were repeated using a confirmatory selection approach. For this approach, all items with  $\beta_i$ , the angle between the item's direction of best measurement and the  $\Theta_1$  axis, greater than  $45^\circ$  were placed in the AT Subtest.

Four levels of the number of examinees were used in the study (750, 1000, 1500, 2000). When selecting the AT subtest using FAC, the first (250, 350, 500, 750) examinees were used to determine the AT Subtest and the remaining (500, 650, 1000, 1250) examinees were used to calculate the DIMTEST statistic. For the confirmatory approach to selecting the AT Subtest, the first (250, 350, 500, 750) examinees were ignored and the remaining (500, 650, 1000, 1250) examinees were used to calculate the DIMTEST statistic. This was done so that the power estimates of the actual DIMTEST run would be conducted using the same sample size. If an actual table of specifications is available, then the entire set of examinees could be used for the DIMTEST run resulting in higher power rates for the procedure.

For all DIMTEST runs, the minimum cell size used in calculating the DIMTEST statistic was set to two (step 2) and the number of resamplings of the data was set to  $N = 50$  (step 8).

All levels of the design: 3 two-dimensional models, 3 tests, 4 examinee levels, 2 AT Selection methods, and 2 examinee ability correlations; were fully crossed, giving 144 different two-dimensional models. Each model was simulated 100 times and the number of rejections of the null hypothesis of unidimensionality recorded in Table 2 for the simple structure model, in Table 3 for the approximate simple structure model, and in Table 4 for the no structure model. The nominal rate of the rejection is  $\alpha = 0.05$ .

Table 2: DIMTEST: Power Results, Simple Structure Model

	Test	ASVAB		ACTM		SATV	
$\rho$	AT method	FAC	CONF	FAC	CONF	FAC	CONF
0.3	J = 750	100	100	100	100	100	100
	J = 1000	100	100	100	100	100	100
	J = 1500	100	100	100	100	100	100
	J = 2000	100	100	100	100	100	100
0.7	J = 750	94	97	95	99	100	100
	J = 1000	97	99	98	99	100	100
	J = 1500	97	99	99	100	99	100
	J = 2000	100	97	97	100	100	100

Table 3: DIMTEST: Power Results, Approximate Simple Structure Model

	Test	ASVAB		ACTM		SATV	
$\rho$	AT method	FAC	CONF	FAC	CONF	FAC	CONF
0.3	J = 750	89	98	100	100	100	100
	J = 1000	93	99	100	100	100	100
	J = 1500	99	100	100	100	100	100
	J = 2000	99	99	100	99	100	100
0.7	J = 750	18	84	61	100	83	100
	J = 1000	17	97	71	99	91	100
	J = 1500	32	100	77	100	93	100
	J = 2000	36	96	84	99	94	100

Table 4: Power Results, No Structure Model

	Test	ASVAB		ACTM		SATV	
$\rho$	AT method	FAC	CONF	FAC	CONF	FAC	CONF
0.3	J = 750	61	98	75	100	100	100
	J = 1000	76	98	84	100	100	100
	J = 1500	90	98	90	100	100	100
	J = 2000	98	99	97	100	100	100
0.7	J = 750	5	45	5	89	24	100
	J = 1000	20	58	5	91	33	100
	J = 1500	26	78	9	100	37	100
	J = 2000	27	80	6	100	34	100

From the simulation study in Froelich (2001), the new DIMTEST procedure has Type I error rates near the nominal rate of  $\alpha = 0.05$ . The power of the new DIMTEST procedure is at or near 100% for all 48 simple structure models and for the 24 approximate simple structure models with low correlation between examinee abilities ( $\rho = 0.3$ ).

For the other three models, approximate simple structure with high correlation between examinee abilities ( $\rho = 0.7$ ), and the no structure model, the power of the DIMTEST procedure using the AT selection method FAC is significantly below 100%, sometimes dramatically so. However, when the DIMTEST procedure is given the 'optimal' AT subtest, the power rates of the DIMTEST procedure are very high for almost all the multidimensional models. The only cases of truly marginal power are for the no structure model with 25 test items and high correlation between examinee abilities ( $\rho = 0.7$ ).

These simulation results showed the need to develop an alternative AT selection method for the new DIMTEST procedure. The goal of the new AT selection method then is to produce a working DIMTEST procedure with similar Type I error levels, but improved power to detect multidimensionality present in the data.

### 3 New AT Selection Method

The proposed AT selection method is based on a combination of the HCA/CCPROX and DETECT procedures. Both procedures are based on the item pair conditional covariance at the heart of the DIMTEST statistic. The underlying mechanism of the DETECT procedure is based on the theory of generalized compensatory models (Zhang & Stout, 1999b). The DETECT statistic depends on both the matrix of examinee response to the items, and on a partitioning of the items into disjoint clusters. The formula for the DETECT statistic for a partition of the test items into cluster is based on the value

$$\frac{2}{n(n-1)} \sum_{1 \leq i < l \leq n} \delta_{i,l} \sum_{k=0}^{n-2} \frac{N_k}{N} \widehat{cov}(U_i, U_l | S_{i,l} = k) \quad (9)$$

where  $N$  is the total number of examinees,  $S_{i,l}$  is the total score on the remaining  $n - 2$  items,  $N_k$  is the number of examinees with score  $S_{i,l} = k$ ,  $\widehat{cov}$  is the standard maximum likelihood estimate of the covariance, and  $\delta_{i,l} = 1$  if items  $i$  and  $l$  are in the same cluster, and  $\delta_{i,l} = -1$  if items  $i$  and  $l$  are in different clusters (Zhang & Stout, 1999a). The DETECT statistic is then a function of the sum of the estimated conditional covariance between items that belong to the same cluster (positive quantities if the items truly belong to the same cluster) minus the estimated conditional covariances between items that belong to the different clusters (negative quantities if the items truly belong to different clusters). Thus, the maximum value of the DETECT statistic will correspond to the partitioning of the test items that most closely mimics the actual structure of the data.

The DETECT procedure uses a genetic algorithm to search for this maximizing partition. In an ideal case, the DETECT procedure would return a partition consisting of several clusters; one of these clusters would be chosen as the AT subtest, while the remaining test items would be the PT subtest. However, there are two primary difficulties with this method; the cluster returned by the DETECT procedure might not be of a desirable size, and when the data is unidimensional, DETECT might only return one cluster consisting of all the test items. (Recall, DIMTEST requires two subtests (AT and PT), and that AT contains an appropriately small percentage of the items.

The HCA/CCPROX procedure uses a proximity matrix based on the item pair conditional covariances to determine the multidimensional structure of the data. The proximity between two items ( $U_i, U_l$ ) is calculated using the formula

$$p_{ccov}(U_i, U_l) = \frac{-1}{N} \sum_{k=0}^{n-2} N_k \widehat{cov}(U_i, U_l | S_{i,l} = k) + \text{constant} \quad (10)$$

where  $N$ ,  $N_k$ ,  $S_{i,l}$ , and  $\widehat{cov}$  are defined as above. (Roussos, et al., 1998). HCA/CCPROX procedure begins with each item in its own separate cluster. At each stage, the procedure joins two clusters together based on the proximity matrix, with the final stage consisting of one cluster of all the test items. HCA/CCPROX differs from the DETECT procedure in that it gives no information on how many clusters are appropriate for the data. In addition, it is also locked into its hierarchical framework, and thus may not find the optimal clustering of items.

To overcome the limitations of both the DETECT and HCA/CCPROX procedure in selecting the AT subtest, the proposed method uses the DETECT statistic to evaluate all possible item partitionings obtained from the HCA/CCPROX procedure. The set of all possible item partitionings is formed by using all clusters from HCA/CCPROX with size between four and one-half of the test items. Each cluster from HCA/CCPROX becomes a possible choice for the AT subtest, while the remaining items become the corresponding PT subtest. Then, these two cluster partitionings of the items are evaluated using the DETECT statistic. The partition with the largest value of the DETECT statistic is chosen as the AT and PT subtests for the DIMTEST procedure.

## 4 Comparative Simulation Study

The simulation study from Froelich (2001) was repeated to determine if the new method based on HCA/CCPROX and DETECT would produce similar Type I error levels as the old FAC program while improving the power of the DIMTEST procedure to detect multidimensionality present in the data. For the Type I error study, all components of the Froelich (2001) study were duplicated. Each unidimensional model was simulated 100 times and the number of rejections of the null hypothesis of unidimensionality recorded in Table 5. (The numbers in the parenthesis are the corresponding results using FAC from Table 1). The nominal rate of rejection is  $\alpha = 0.05$ .

Table 5: DIMTEST: Type I Error Results, New AT Method

Test	ASVAB	ACTM	SATV
J = 750	5 (1.50)	2 (2.25)	0 (2.00)
J = 1000	3 (3.00)	1 (4.50)	0 (1.75)
J = 1500	6 (4.00)	4 (4.75)	0 (2.75)
J = 2000	11 (2.00)	7 (3.00)	2 (3.50)

For the power study, all components of the Froelich (2001) study were duplicated. Each multidimensional model was simulated 100 times and the number of rejections of the null hypothesis of unidimensionality recorded in Table 6 for the simple structure model, in Table 7 for the approximate simple structure model, and in Table 8 for the no structure model. (Again, the numbers in parenthesis are the corresponding results using FAC from Tables 2, 3, and 4.

Using the new AT selection method, the DIMTEST procedure has Type I error rates at or near the nominal rate of  $\alpha = 0.05$  for all but one unidimensional model. The Type I error rate is slightly inflated for the ASVAB 25 item test with 2000 examinees. A later simulation will show this inflation is reduced by using a larger percentage of examinees to select the AT subtest.

Table 6: DIMTEST: Power Results, Simple Structure Model, New AT Method

$\rho$	Test	ASVAB	ACTM	SATV
0.3	J = 750	100 (100)	100 (100)	100 (100)
	J = 1000	100 (100)	100 (100)	100 (100)
	J = 1500	100 (100)	100 (100)	100 (100)
	J = 2000	100 (100)	100 (100)	100 (100)
0.7	J = 750	95 (94)	99 (95)	98 (100)
	J = 1000	99 (97)	100 (98)	100 (100)
	J = 1500	98 (97)	100 (99)	100 (99)
	J = 2000	96 (100)	100 (97)	100 (100)

Table 7: DIMTEST: Power Results, Approximate Simple Structure Model, New AT Method

$\rho$	Test	ASVAB	ACTM	SATV
0.3	J = 750	100 (89)	100 (100)	100 (100)
	J = 1000	96 (93)	100 (100)	100 (100)
	J = 1500	99 (99)	100 (100)	100 (100)
	J = 2000	100 (99)	99 (100)	100 (100)
0.7	J = 750	38 (18)	85 (61)	100 (83)
	J = 1000	46 (17)	90 (71)	100 (91)
	J = 1500	67 (32)	98 (77)	100 (93)
	J = 2000	90 (36)	97 (84)	100 (94)

Table 8: DIMTEST: Power Results, No Structure Model, New AT Method

$\rho$	Test	ASVAB	ACTM	SATV
0.3	J = 750	46 (61)	76 (75)	100 (100)
	J = 1000	74 (76)	93 (84)	100 (100)
	J = 1500	86 (90)	99 (90)	100 (100)
	J = 2000	96 (98)	98 (97)	100 (100)
0.7	J = 750	9 (5)	14 (5)	56 (24)
	J = 1000	12 (20)	21 (5)	81 (33)
	J = 1500	16 (26)	32 (9)	99 (37)
	J = 2000	32 (27)	59 (6)	100 (34)

Using the new AT selection method, the DIMTEST procedure still produces power rates at or near 100% for the simple structure models and the approximate simple structure model with low correlation ( $\rho = 0.3$ ). For the approximate simple structure model with high correlation, an increase in the power of the DIMTEST procedure with the new AT selection method is present across all 12 item and examinee levels, with an average increase from 63.08 with FAC to 84.25 with the new AT selection method.

For the no structure models, the power of the DIMTEST procedure with the new AT selection method is at or above the power of the DIMTEST procedure using the FAC program. There are two dramatic increases in power in the no structure models with high correlation ( $\rho = 0.7$ ) and either 40 or 80 test items. For the 40 test items, there is an average increase from 6.25 with FAC to 31.5 with the new AT selection method and for the 80 test items, there is an average increase from 32 with FAC to 84 with the new AT selection method.

The simulation study shows the new AT selection method gives the DIMTEST procedure a Type I error rate at or near the nominal rate of  $\alpha = 0.05$ . In addition, the new AT selection method produces a DIMTEST procedure with at least similar, and in some cases, greatly improved power to detect multidimensionality present in the data. Thus, the new AT selection method based on HCA/CCPROX and DETECT makes better use of the potential of the DIMTEST statistic for detecting multidimensionality present in the data. The new AT selection method therefore produces an overall improvement in the DIMTEST procedure.

#### 4.1 Additional Simulation Studies

The results of the power simulation study for the DIMTEST procedure with the new AT selection method were an overall improvement over the simulation results using the FAC selection method for AT. However, the simulation results when compared to the confirmatory approach for selecting AT show room for improvement in the power of the DIMTEST procedure. For the approximate simple structure model with high correlation ( $\rho = 0.7$ ) and the no structure model, the power of the DIMTEST procedure still fails to approach the levels of power for the 'optimal' choice of AT, especially with a small number of items and a high correlation ( $\rho = 0.7$ ).

One possible source of improvement in the power rates of the new DIMTEST procedure is the division of examinees between selecting the AT subtest and calculating the DIMTEST statistic. The study in Froelich (2001) and the study on the new AT selection method used approximately one-third of the examinees to select the AT subtest. The study below was then conducted to determine if increasing the percentage of examinees used to select the AT subtest would improve the power of the DIMTEST procedure.

There is generally a trade-off in Type I error rates and power rates when dividing the examinee responses between AT subtest selection and DIMTEST calculation. When the data are unidimensional, the AT subtest is in effect randomly chosen. Using more examinees for selecting the AT subtest will not result in better Type I error rates. In fact, Type I error rates could suffer as a result of having fewer examinees present to calculate the DIMTEST statistic. However, when the data are multidimensional, using more examinees for selecting the AT subtest will hopefully result in the selection of a better AT subtest. This in turn will improve the power of the DIMTEST procedure. Thus the examinees should be divided in such a way as to maintain good Type I error rates but increase the power of the procedure.

For this simulation study, three different levels for the division of examinees between AT selection and DIMTEST calculation were studied; 33% AT selection: 67% DIMTEST calculation, 50% AT selection: 50% DIMTEST calculation, and 67% AT selection: 33% DIMTEST calculation. The design, parameters, models, etc. of the study were the same as in Froelich (2001). For the Type

I error study, each unidimensional model was simulated 100 times and the number of rejections of the null hypothesis recorded in Table 9. The nominal rate of rejection is  $\alpha = 0.05$ .

Table 9: DIMTEST: Type I Error Results, New AT Method

TEST	ASVAB			ACTM			SATV		
% for AT	33	50	67	33	50	67	33	50	67
J = 750	5	3	0	2	2	2	0	3	2
J = 1000	3	1	2	1	2	1	0	2	0
J = 1500	6	6	3	4	0	3	0	4	2
J = 2000	11	7	6	7	6	1	2	4	2

For the power study, only three multidimensional models were simulated, the approximate simple structure model with high correlation ( $\rho = 0.7$ ) and the no structure model. Each multidimensional model was simulated 100 times and the number of rejections of the null hypothesis of unidimensionality recorded in Table 10 for the approximate simple structure model and Table 11 for the no structure model. The nominal rate of rejection is  $\alpha = 0.05$ .

Table 10: DIMTEST: Power Results, Approximate Simple Structure Model, New AT Method

TEST	ASVAB			ACTM			SATV		
% for AT	33	50	67	33	50	67	33	50	67
J = 750	38	39	41	85	89	90	100	100	100
J = 1000	46	59	63	90	96	98	100	100	100
J = 1500	67	81	78	98	99	97	100	100	100
J = 2000	90	86	92	97	97	97	100	100	100

Table 11: DIMTEST: Power Results, No Structure Model, New AT Method

	TEST	ASVAB			ACTM			SATV		
$\rho$	% for AT	33	50	67	33	50	67	33	50	67
0.3	J = 750	46	65	53	76	85	87	100	100	100
	J = 1000	74	69	68	93	91	93	100	100	100
	J = 1500	86	91	86	99	98	98	100	100	100
	J = 2000	96	96	93	98	98	98	100	100	100
0.7	J = 750	9	9	4	14	11	11	56	81	80
	J = 1000	12	14	11	21	21	18	81	87	89
	J = 1500	16	27	21	32	36	41	99	100	98
	J = 2000	32	32	27	59	66	56	100	100	98

The simulation results show the DIMTEST procedure still has Type I error rates at or near the nominal rate when either 50% or 67% of the examinees are used to select the AT subtest. In fact, the Type I error rates appear to be closer to the nominal with no inflation present when more

examinees are used to select the AT subtest. The reason for this result is not clear at this time, but is under study.

The simulation results also show an improvement in the power of the DIMTEST procedure when 50% of the examinees are used to select the AT subtest. However, there is not an overall improvement in the power rates for the DIMTEST procedure when the percentage of examinees used to select the AT subtest was 67%. Thus, we recommend, when using the new AT selection method, dividing the examinees into two equal groups, with the first group used to determine the AT subtest, and the second group used to calculate the DIMTEST statistic.

## 5 Conclusion and Future Research

The original DIMTEST procedure (Stout, 1987) has been shown to be among the most powerful of the hypothesis testing procedures for testing the assumption of unidimensionality for large scale tests. Recently, a new DIMTEST procedure was developed (Froelich, 2001) that removed the need for the AT2 subtest of items. A simulation study in Froelich (2001) showed this new version of DIMTEST has Type I error rates near the nominal rate and good power to detect multidimensionality in several models. However, an analysis of the power of the procedure showed using linear factor analysis to select the AT subtest resulted in very low power rates for some multidimensional models.

The goal of this research then was to replace the linear factor analysis program (FAC) associated with DIMTEST with a method based on the item pair conditional covariances used by the DIMTEST statistic. The new method of AT selection is a combination of the programs HCA/CCPROX and DETECT. With this new selection method, the DIMTEST procedure still has Type I error rates near the nominal rate, but has at least similar, and in most cases, greatly improved power rates to detect multidimensionality for the simulated multidimensional models.

Although the power rates for the new AT selection method were higher, there was still room for improvement. A second simulation study was conducted to determine if increasing the percentage of examinees used to select the AT subtest would result in increased power for the DIMTEST procedure. The simulation study showed the power of the DIMTEST procedure increased when 50% of the examinees were used to select the AT subtest. However, the power of the DIMTEST procedure showed no significant changes when 67% of the examinees were used to select the AT subtest.

There are several areas of research left to be studied before a final version of the AT selection method is reached. First, changes to the method should be studied in an attempt to further improve the power of the DIMTEST procedure. These changes include using a slightly different DETECT statistic to improve the selection of AT. Second, additions to the simulation study should be made to determine the power of the DIMTEST procedure under different unidimensional and multidimensional models. These additions include using different item parameter sets, changing the percentage of items that best measure  $\theta_1$ , changing the examinee ability distribution to a non-normal distribution, and changing from a compensatory to a non-compensatory multidimensional model.

## References

- Drasgow, F.** (1987). A study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72*, 19-30.
- Froelich, A.G.** (2001) A new bias correction method for DIMTEST. Unpublished manuscript.
- Hattie, J., Krakowski, K., Rogers, J., & Swaminathan, H.** (1996) An assessment of Stout's index of essential dimensionality. *Applied Psychological Measurement, 20*, 1-14.
- Hulin, C.L., Drasgow, F., & Parsons, L.K.** (1983). *Item response theory*. Homewood, IL: Dow Jones-Irwin.
- 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.
- McDonald, R.P** (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34*, 100-117.
- 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 multidimensional model for dichotomous items response data*. In W.J. van der Linden & Hambleton, R.K. (Eds), *Handbook of Modern Item Response Theory*. New York:Springer.
- Roussos, L.A., Stout, W., & Marden, J.** (1998) Using new proximity measures with heirarchical cluster analysis to detect multidimensionality. *Journal of Educational Measurement, 35*, 1-30.
- Stout, W.** (1987) A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589-617.
- Zhang, J. & Stout, W.** (1999a). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64*, 213-249.
- Zhang, J. & Stout, W.** (1999b). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika, 64*, 129-154.