DOCUMENT RESUME

ED 371 011 TM 021 642

AUTHOR Nandakumar, Ratna; Yu, Feng
TITLE Testing the Robustness of DIMTEST on Nonnormal Ability Distributions.
PUB DATE Apr 94
NOTE 16p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 5-7, 1994).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Ability; Content Validity; Correlation; Nonparametric Statistics; Probability; *Responses; *Robustness (Statistics); Sample Size; Simulation; *Statistical Distributions; Statistical Studies; *Test Items; Test Length; Test Validity
IDENTIFIERS *DIMTEST (Computer Program); Item Characteristic Function; Nonnormal Distributions; Stouts Procedure; *Unidimensionality (Tests)

ABSTRACT

DIMTEST is a statistical test procedure for assessing essential unidimensionality of binary test item responses. The test statistic T used for testing the null hypothesis of essential unidimensionality is a nonparametric statistic. That is, there is no particular parametric distribution assumed for the underlying ability distribution or for the item characteristic curves generating item response in the mathematical derivation of probability distribution of the statistic T. The purpose of the present study is to empirically investigate the robustness of the statistic T with respect to ability distributions. Several nonnormal distributions, both symmetric and nonsymmetric, are considered in simulations involving six different types of ability distributions. In addition, test length and sample size are used as parameters in the present study. Simulation results indicate that the performance of Stout's statistics T subscript c and T subscript p are consistent with their theoretical developments, in that no particular shape is assumed for examinee abilities. That is, these statistics are robust against the shape of the ability distribution. Included are seven tables. (Contains 8 references.) (Author/SLD)

Testing the Robustness of DIMTEST on Nonnormal Ability Distributions

Ratna Nandakumar
Feng Yu
Department of Educational Studies
University of Delaware

Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, April 6, 1994

Testing the Robustness of DIMTEST on Nonnormal Ability Distributions

Abstract

DIMTEST is a statistical test procedure for assessing essential unidimensionality of binary item responses. The test statistic T used for testing the null hypothesis of essential unidimensionality is a nonparametric statistic. That is, there is no particular parametric distribution assumed for the underlying ability distribution or for the item characteristic curves generating item responses in the mathematical derivation of probability distribution of the statistic T. The purpose of the present study is to empirically investigate the robustness of the statistic T with respect to ability distributions. Several nonnormal distributions, both symmetric and nonsymmetric, are considered for this purpose. In addition, test length and sample size are used as parameters in the present study.

Currently, unidimensional IRT models are most commonly used models for drawing inferences about an examinee's standing on a trait of interest based on his/her responses to a set of items. Assessment of unidimensionality of item response data is therefore essential prior to applying any unidimensional model to data so that meaningful inferences can be drawn about the examinee's relative standing on the trait of interest. DIMTEST is a statistical procedure to assess for unidimensionality of binary item response data. It was first developed by Stout (1987) and subsequently refined by Nandakumar and Stout (1993). DIMTEST was developed to assess whether a given data of item responses fit an essentially unidimensional model. That is, it assesses if there is one dominant ability (proficiency) driving item responses. DIMTEST is a nonparametric test, which means that there is no particular parametric distribution assumed for the underlying abilities or for the type of item characteristic curves (ICCs) generating the item responses in the mathematical derivation of probability distribution of the test statistic T. The only asumptions made in the development of DIMTEST methodology are: a) essential independence, b) random sample of examinees from a population, and c) monotonitically increasing item response functions.

The assumption of essential independence is crucial for the DIMTEST procedure to work well. The assumption of essential independence requires that conditional item responses be independent of one another when conditioned upon the dominant ability. This is a weaker form of the assumption of local independence which requires conditioning on all abilities, major and minor, influencing item responses rather than just the major ability.

In all simulation studies conducted so far (Nandakumar, 1991; Nandakumar and Stout, 1993; Stout 1987) the examinee abilities were generated from the standard normal distribution. In practical applications, however, it is possible to observe nonnormal ability distributions. Therefore, it is important to know how DIMTEST performs in these situations. Athough on theoretical basis it should perform well, we need to establish this empirically. The purpose of the present study is therefore to do a detailed and extensive investigation of robustness of DIMTEST on various nonnormal ability distributions that may underlie item responses. Six different ability distributions were considered with varying levels of test lengths and sample sizes. For each case of ability distributions it was of interest to note if the observed level of

significance matches that of the nominal level across sample sizes and test lengths. Also, of interest was to observe if the distribution of Ts follows a normal distribution for different ability distributions.

In order to eliminate the indeterminacy that exists between the ability distribution and the functional forms of the item characteristic curves (ICC), throughout, we fix the ICCs to be of logistic form while varying the ability distributions.

DIMTEST Procedure

The hypothesis to test for essential unidimensionality can be stated as

$$H_0 : d_E = 1 \; vs. \; H_1 : d_E > 1$$

where d_E denotes essential dimensionality.

DIMTEST assesses unidimensionality of a given set of item responses (found as a result of administering a set of items to a group of examinees) by splitting the items into three subtests: two short subtests AT1 and AT2, and a large subtest PT. Items of AT1 are first selected so that they are dimensionally homogeneous. Items of AT2 are matched in difficulty to the items of AT1 and the rest of the items form the subtest PT. There are several methods to select AT1 items. Simple factor analysis can be used to select AT1 items. Using factor analysis, a small set of items with highest loadings of the same sign on the second factor are selected (Nandakumar & Stout, 1993; Stout, 1987). Expert opinion is another method to select AT1 items. Based on experience, one can select a small set of items (not more than one-quarter of total items) believed to be measuring the same trait (Nandakumar, 1993). Alternatively, one can use hierarchical cluster analysis to select AT1 items (Roussos, Stout, and Marden, 1993). In the present study factor analysis was used to select AT1 items.

If item responses were driven by an essentially unidimensional model, items of all the subtests (AT1, AT2, and PT) would be of similar dimension. On the contrary, if item responses were driven by a multidimensional model,

items of AT1 will be dimensionally homogeneous and differ from the rest of the items in the dimensional structure.

Item responses of the subtest PT are used to group examinees into K subgroups. Item responses of the subtest AT1 are u- d to compute two variance estimates $\hat{\sigma}_k^2$ and $\hat{\sigma}_{U,k}^2$ within each subgroup and their difference is appropriately standardized and summed across subgroups to arrive at the statistic T_1 given by

$$T_1 = \frac{1}{\sqrt{K}} \sum_{k=1}^{K} \left[\frac{\hat{\sigma}_k^2 - \hat{\sigma}_{U,k}^2}{S_k} \right]$$

where $\hat{\sigma}_k^2$ is the variance estimate of the AT1 subtest among examinees in the subgroup k, and $\hat{\sigma}_{U,k}^2$ is the estimate of the "unidimensional" variance computed by summing the item varinances of the subtest AT1. S_k is the appropriate standard error computed for the subgroup k (for details see Nandakumar & Stout, 1993).

Similarly item responses of AT2 are used to compute variance estimates within subgroups and their difference is standardized and summed across subgroups to arrive at the statistic T_2. Stout's statistic T to assess for essential unidimensionality is then given by

$$T = \frac{T_1 - T_2}{\sqrt{2}}$$

which follows the standard normal distribution when the null hypothesis of essential unidimensionality holds.

When the null hypothesis of essential unidimensionality holds, the two variance estimates $\hat{\sigma}_k^2$ and $\hat{\sigma}_{U,k}^2$ should be approximately equal resulting in a small value for T_1. Although this is generally true, in certain situations T_1 could be inflated under H_0 due to difficulty nature of items in AT1 or due to shortness of the subtest PT (Nandakumar, & Stout, 1993; Stout, 1987). In such situations the statistic T_2 was designed to correct T_1 for inflation due to statistical biases. Consequently, under H_0, T will be small leading to the tenability of H_0. On the other hand when H_1 holds, the difference in the variance estimates will be large leading to the rejection of H_0.

The performance of DIMTEST has been studied extensively through Monte Carlo simulations in various test settings by varying parameters such as test length, sample size, ICC type, correlation between abilities, and the degree of multidimensionality (Nandakumar and Stout, 1993; Nandakumar, 1991; Stout, 1987). It has also been studied for its performance on various real tests (Nandakumar, 1993). It has been found that DIMTEST has maintained desirable type-I error with high power even when the correlation between abilities is as large as 0.7.

Method

In order to study the robustness of the statistic T due to different ability distributions, a modest size simulation study was designed with three factors varied: the type of ability distribution, test length, and sample size. Six different types of ability distributions were considered to generate examinee abilities— normal distribution, bimodal distribution, positively skewed distribution, negatively skewed distribution, "positive" chi-square distribution, and "negative" chi-square distribution.

Bimodal distribution was chosen to represent a situation where two radically different types of examinees take the test. The bimodal distribution was formed as a mixture of two normal distributions with equal probabilities for each component in the mixture. The means of each of the components were -1.5 and 1.5. The mean of the bimodal distribution is 0 and the variance is 3.25.

The positively skewed distribution was chosen to represent a situation where most examinees are of low to moderte ability. The mean of this distribution is 0 and the variance is 1. The positively skewed distribution was generated using the power method suggested by Fleishman (1978). Using this approach a transformation is applied to a random variable from a normal distribution as follows:

$$Y = a + bX + cX^2 + dX^3$$

where X follows the standard normal distribution and the constants a, b, c,

and d are the weights. Fleishman (1978) lists in a table these weights for different values of skewness and kurtosis. In our study we took appropriate weights to have a skewness of 0.75 and a kurtosis of 0.5. These values of skewness and kurtosis were chosen because it is believed, based on empirical observaions, that a "typical" nonnormality occurs with skewness less than 0.8 and kurtosis between -0.6 and 0.6 (Fleishman, 1978).

The negatively skewed distribution was chosen to represent a situation where most examinees are of moderate to high ability. The mean of this distribution is 0 and the variance is 1. This distribution was generated similar to the positively skewed distribution with appropriate weights for the negatively skewed distribution to reflect a skewness of -0.75 and a kurtosis of 0.5.

The "positive" chi-square distribution was generated by linearly transforming a chi-square random variable as follows:

$$Y = (X - 6) * 0.55$$

where X follows the chi-square distribution with six degrees of freedom. After a trial and error process a chi-square distribution with six degrees of freedom was chosen so that we have a reasonable distribution of abilities. The other choices of degrees of freedom either had too little variance or too much variance and was resulting in too many 0s or 1s in the response pattern. A multiplicative factor of 0.55 was chosen in order to make the mode approximately equal to -1 (the exact mode is -1.1). In this way this distribution represents a situation where most examinees are of low ability. This distribution has a mean of 0 and a variance of 3.63. The "negative" chi- square distribution was obtained by changing the sign of the positive chi-square distribution to represent a situation where most examinees are of high ability. The mode of this distribution is 1.1. The mean of the negative chi-square distribution is 0 and the variance is 3.63. The main difference between the skewed and the chi-square distributions is the variability among examinee abilities. For the chi-square distributions higher proportion of examinees fall into the extreme groups than the skewed distributions. Table 1 lists all the distributions with their means and variances.

Four different sample sizes were considered: 500, 750, 1000, and 1500. Two test lengths were considered: 20 and 50. These test lengths were considered to represent a typical "short" test and a typical "long" test. Examinee responses were generated using the three-parameter logistic unidimensional model given by

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + exp(-1.7a_i(\theta - b_i))}$$

The item parameters of the SAT-verbal test were obtained from the literature (Lord, 1968) in order to make the simulations as realistic as possible. The SAT-verbal test consists of 80 items. The desired number of items were randomly selected from the 80-item pool. For a given theta level and given item, the probability of correct response was obtained using the above three-parameter logistic model. A random number between 0 and 1 was generated from a uniform distribution. If the computed probability, $P_i(\theta)$ was greater than or equal to the random number generated, the examinee was said to have answered the item correctly and was given a score of 1; otherwise the examinee was given a score of 0.

For each combination of theta distribution, test length, and sample size, examinee responses were analyzed for unidimensionality using DIMTEST. As explained before DIMTEST uses part of the sample for factor analysis to select subtest items (if this method is chosen) and the rest of the sample is used to compute Stout's statistic T. In previous studies it was shown that a minimum sample of 250 is needed for factor analysis and a sample of 500 is optimum. Therefore, in the present study a subsample of 250 examinees was randomly selected from the given sample for factor analysis when the sample size is either 500 or 750. And a subsample of 500 examinees was selected for factor analysis when the given sample size was 1000 or 1500. In Tables 2-7, second and third columns denote the sample sizes used for the factor analysis (J_F) and for computing the statistic (J_S) for each simulation.

Examinee responses for each combination of factors were simulated 100 times. Each time the responses were assessed for unidimensionality with the nominal type-I error rate set to 0.05. The number of rejections for 100 replications was observed.

Results

The observed type-I error rates (rejection rate) for all ability distribution types are listed in Tables 2-7. Each table contains results for one type of ability distribution for all combinations of test lengths and sample sizes. For each such combination, rejection rates for two values of T: T_c and T_p are reported. The statistic T_c denotes the conservative statistic and the statistic T_p denotes the more powerful statistic. T_c is the statistic that was originally developed by Stout (1987) and was found to be slightly conservative. That is, in simulation studies it exhibited, on the average, slightly lower observed type-I error rate than the nominal error rate. The latter refined statistic T_p (Nandakumar & Stout, 1993) has been found to be exhibiting an observed type-I error rate that is closer to the nominal error rate with much higher power than T_c. In previous simulation studies the T_p was also found to be slightly inflated when small sample sizes (such as less than 1000) are associated with long tests (such as 40 or 50). DIMTEST reports both these statistics.

Looking through the results in Tables 2-7 it can be seen that the observed type-I error rate for both T_c and T_p is less than or equal to the nominal level of .05 for all distribution types and sample sizes when the test length is 20. For the case of test lenght of 50, however, the observed type-I error rate is within the nominal level only for the conservative statistic T_c for all distribution types and sample sizes. For the more powerful statistic T_p, there is a slight inflation in some cases, especially when the sample size is less than 1000 with test length of 50. The most important result, however, is that the results are consistent across all different distribution types for both statistics T_c and T_p, supporting the conclusion that both are nonparametric statistics and hence robust against different ability distributions.

A random sample of these runs were selected and tested to see if the distribution of Ts across 100 replications follows the unit normal distribution. The results were positive confirming normal distribution for Ts.

Summary and Conclusions

In summary, simulation results in the present study indicate that the performance of Stout's statistics T_c and T_p are consistent with their theoretical developments, in the sense that no particular shape is assumed for examinee abilities. That is, these statistics are robust against the shape of the abilty distribution. In the present study, both statistics T_c and T_p have in general shown good adherence to the nominal level across all distributions. As recommended previously (Nandakumar, & Stout, 1993), T_p is the recommended statistic to use generally. In cases of small samples associated with long tests (such as 50 items with less than 1000 examinees as seen in the present study), however, it is advisable to look into both statistics.

The results obtained in this study were compared to those in De Champlain and Tang's (1993) study where they looked at robustness of Stout's statistic T and two other chi-squared based statistics to assesss unidimensionality on three theta distributions (normal, positively skewed, and negatively skewed). In their study positive skewness was set to 0.75, same as the present study. However, the negative skewness in their study was set to -1.25, which is considered as not "typical" according to empirical studies of nonnormal distributions (Fleishman, 1978). Their study used test lengths of 20 and 40 and sample sizes of 500 and 1000. De Champlain and Tang's study showed inflation of type-I error rate for the statistic T for skewed distributions for both sample sizes and test lengths except for shorter length tests and positively skewed cases. Comparing their results with those obtained here we find that with a skewness of -1.25 all examinees were located in a small ability range resulting in not having enough examinees at certain test score levels leading to inconsistent values for T. Whereas in the case of positive skewed distributions the results are consistent in both studies. That is, larger samples are needed with larger test lengths in order to use the more powerful test statistic T_p.

References

1. De Champlain, A., & Tang, L. (1993, April). *The effect of nonnormal ability distributions on the assessment of dimensionality.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, Georgia.

2. Fleishman, A. (1978). A method for simulating non-normal distributions. *Psychometrika, 43*, 521-532.

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

4. Nandakumar, R. (1991). Traditional dimensionality vs essential dimensionality. *Journal of Educational Measurement,28*, 99-117.

5. Nandakumar, R. (1993). Assessing dimensionality of real data. *Applied Psychological Measurement,17*, 29-38.

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

7. Roussos, L. A., Stout, W. F., & Marden, J. I. (1993, April). *Dimensional and structural analysis of standardized tests using DIMTEST with hierarchical cluster analysis.* Paper presented at the annual National Council on Measurement in Education meeting, Atlanta, Georgia.

8. Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika, 55*, 293-325.

Table 1

Ability distributions and their means and standard deviations

	Ability Distribution					
	Normal	Bimodal	Positively Skewed	Negatively Skewed	"Positive" Chi-square	"Negative" Chi-square
Mean θ	0	0	0	0	0	0
Variance θ	1	3.25	1	1	3.63	3.63
Skewness θ	0	0	0.75	-0.75	–	–

Table 2

Distribution of θ: Normal

Test Length	Sample Size		Rejection Rate	
N	J_F	J_S	T_c	T_p
20	250	250	1	4
20	250	500	1	5
20	500	500	1	5
20	500	1000	1	3
50	250	250	4	6
50	250	500	3	9
50	500	500	3	8
50	500	1000	4	7

Table 3

Distribution of θ: Bimodal

Test Length	Sample Size		Rejection Rate	
N	J_F	J_S	T_c	T_p
20	250	250	1	2
20	250	500	0	4
20	500	500	0	4
20	500	1000	0	3
50	250	250	0	1
50	250	500	1	2
50	500	500	1	4
50	500	1000	2	5

Table 4

Distribution of θ: Positive Skewed

Test Length	Sample Size		Rejection Rate	
N	J_F	J_S	T_c	T_p
20	250	250	1	2
20	250	500	0	0
20	500	500	0	2
20	500	1000	1	3
50	250	250	3	9
50	250	500	1	10
50	500	500	0	5
50	500	1000	2	4

Table 5

Distribution of θ: Negative Skewed

Test Length	Sample Size		Rejection Rate	
N	J_F	J_S	T_c	T_p
20	250	250	0	3
20	250	500	0	3
20	500	500	1	5
20	500	1000	2	7
50	250	250	3	5
50	250	500	3	10
50	500	500	4	6
50	500	1000	2	5

Table 6

Distribution of θ: Positive χ^2

Test Length	Sample Size		Rejection Rate	
N	J_F	J_S	T_c	T_p
20	250	250	1	2
20	250	500	0	0
20	500	500	0	2
20	500	1000	1	3
50	250	250	3	9
50	250	500	1	10
50	500	500	2	4
50	500	1000	0	5

Table 7

Distribution of θ: Negative χ^2

Test Length	Sample Size		Rejection Rate	
N	J_F	J_S	T_c	T_p
20	250	250	3	5
20	250	500	2	2
20	500	500	3	5
20	500	1000	0	2
50	250	250	6	12
50	250	500	4	9
50	500	500	3	8
50	500	1000	2	6