ED 295 955 TM 011 711

AUTHOR Nandakumar, Ratna
TITLE Modification of Stout's Procedure for Assessing Latent Trait Unidimensionality.
PUB DATE Apr 88
NOTE 11p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 5-9, 1988).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Algorithms; *Latent Trait Theory; *Monte Carlo Methods; Standardized Tests; Statistical Bias; *Statistical Significance; Statistical Studies
IDENTIFIERS *Parametric Analysis; *Stouts Procedure

ABSTRACT

The effectiveness of Stout's procedure for assessing latent trait unidimensionality was studied. Strong empirical evidence of the utility of the statistical test in a variety of settings is provided. The procedure was modified to correct for increased bias, and a new algorithm to determine the size of assessment sub-tests was used. The following two issues were addressed via a Monte Carlo simulation: (1) the ability to approximate the nominal level of significance via the observed level of significance; and (2) the power of the statistical test while undergoing the modifications. Results indicate that Stout's statistic and the procedure, which grew directly out of a meaningful conceptual definition of test dimensionality, avoids the issue of parametric model correctness, attracts the support of asymptotic theory, has modest computational requirements, and receives support from Monte Carlo simulations. (TJH)

Modification of Stout's Procedure for Assessing Latent Trait Unidimensionality

Ratna Nandakumar
University of Delaware

Paper presented at the annual meeting of the American Educational Research Association, New Orleans, April, 1988

Setting: Standardized test of N items administered to J examinees. Each item scored right or wrong:

$$[U_i = 1] = [\text{i-th item right}]$$

Each examinee is characterized by unobservable latent ability θ usually assumed continuous and unidimensional random variable.

Assumptions Underlying the Item Response Theory Models:

1. Unidimensionality, d=1

2. p(θ) increases in θ, "monotonicity"

3.$p[\underline{U} = \underline{u}|\theta] = \prod_{i=1}^{N} p[U_i = u_i|\theta]$ for all $\underline{u}$ and θ

Most Commonly Used Item Characteristic Curve

$$p_i(\theta) = p[U_i = 1|\theta] = c_i + \frac{1-c_i}{1+\exp(-1.7a_i(\theta-b_i)}$$

Where a = discriminatory power of the item,
b = location parameter, or difficulty of the item,
and c = lower asymptote or pseudo guessing parameter of the item

The Definition of Unidimensionality:

What is meant by unidimensionality is that only one dominant dimension or attribute influences the test performance. The dominant attribute results when an attribute is common to many items in a test. Stout (Psychometrika, Dec 87) defines unidimensionality in terms of dominant dimension as follows:

A test $(U_1,U_2,\ldots,U_N)$ of length N is said to be essentially unidimensional if there exists a latent variable θ such that for all values of θ,

$$(1.1) \qquad \frac{1}{N(N-1)} \sum_{1\le i\ne j\le N} |\text{cov}(U_i,U_j|\theta)| \cong 0.$$

This is to say that on the average the conditional covariances should be small. Stout furthur defines an <u>empirical notion of unidimensionality</u> consistent with Equation 1.1 as follows:

A test $(U_i, U_2, \ldots U_N)$ is unidimensional if for <u>all</u> subtests $\{(U_{k_1}, U_{K_2}, \ldots U_{K_M})\}$ of length M (<N) and all values of Y_p,

$$(1.2) \qquad S_{M,N} = \frac{1}{M(M-1)} \sum_{1 \le i \ne j \le M} \mathrm{cov}(U_{k_i}, U_{k_j} | Y_p) \cong 0$$

where Y_p is the proportion correct on the long subtest complementary to $(U_{k_1}, U_{k_2}, \ldots, U_{k_M})$ with length $n = N-M$.

The above definition suggests splitting the items into two subtests such that failure of Equation 1.2 is evidence of lack of unidimensionality.

<u>Stout's Test for Assessing Unidimensionality</u>

$$H_0: d = 1 \quad \text{vs} \quad H_1: d > 1$$

<u>The main underlying assumptions</u>

1. Examinees are randomly selected from a large population.
2. Examinees respond to items independetley of one another.
3. Monotonicity.
4. Local independence (LI): for examinees of same ability, their responses to different items are independent.

Stout's procedure is nonparametric, but 3PL was used in simulation studies.

<u>Test Procedure</u>

<u>Step1</u>. Split N test items into 3 subtests: two short assessment subtests of length M each, and a long partitioning subtest of length n (=N-2M).

AT1—Assessment 1 subtest, length M. Choose items into this subtest so that they are as unidimensional as possible. This can be done either on the basis of expert opinion or using factor analysis as a data analytic tool. The <u>purpose of AT1</u> is to compute Stout's statistic.

AT2—Assessment 2 subtest, length M. After selecting items into AT1, items into AT2 are selected so that they have the same difficulty distribution as those in AT1. The purpose of AT2 is to correct for the pre-asymptotic statistical bias in Stout's statistic.

PT—Partitioning subtest, length n. After selecting items into AT1 and AT2, the rest of the n = N-2M items are put in PT. The purpose of PT is to group examinees into subgroups. When d = 1 and the test is long, each subgroup will consist of examinees approximately of equal ability.

Example 1: d = 1, N = 30, say all verbal items.

AT1 = 5v; AT2 = 5v; PT = 20V

Example 2: d = 2, N = 30, say 10 math (M), 10 verbal (V), and 10 mixed (X) items.

AT1 = 5v; AT2 = 1V, 2M, 2X; PT = 4V, 8M, 8X.

Step 2: Assign examinees into different subgroups according to their scores on PT.

Step 3: Within each examinee subgroup, estimate examinee variation on AT1 in two ways:

Let U_{ijk} denote the response of the j-th examinee to the i-th item from subgroup k. Let J_k denote the number of examinees in subgroup k and let K denote the total number of subgroups.

$$\hat{\sigma}_k^2 = \Sigma_{j=1}^{J_k} (Y_j^{(k)} - \bar{Y}^{(k)})^2 / J_{k},$$

is the usual variance estimate for subgroup k, where $Y_j^{(k)} = \Sigma_{i=1}^{M} U_{ijk}/M$,

and $\bar{Y}^{(k)} = \Sigma_{j=1}^{J_K} Y_j^{(k)}/J_K$, and

$$\hat{\sigma}_{U;k}^2 = \Sigma_{i=1}^{M} \hat{p}_i^{(k)}(1-\hat{p}_i^{(k)})/M^2,$$

is the unidimensional variance for subgroup k, where $p_i^{(k)} = \Sigma_{j=1}^{J_k} U_{ijk}/J_k$.

Note: Under H_0, if the test is long, the long partitioning test ensures that examinees within each subgroup are approximately equal ability and the assumption of local independence will be closely met leading to $\hat{\sigma}_k^2 \cong \hat{\sigma}_{U,k}^2$. Under H_1, however, the assumption of local independence will be badly violated leading to $\hat{\sigma}_k^2 >> \hat{\sigma}_{U,k}^2$.

Step4: Compute the statistic T_L (L for long test) for items in AT1 subtest:

$$(1.3) \qquad T_L = \frac{1}{\sqrt{k}} \Sigma_{k=1}^{K} \left[\frac{\hat{\sigma}_k^2 - \hat{\sigma}_{U,k}^2}{S_k} \right]$$

where S_k is the appropriate normalizing constant given by:

$$S_k^2 = \left[(\hat{\mu}_{4,k} - \hat{\sigma}_k^4) + \hat{\delta}_{4,k}/M^4 + 2\sqrt{(\hat{\mu}_{4,k} - \hat{\sigma}_k^4)\, \hat{\delta}_{4,k}/M^4} \right] / J_k$$

where

$$\hat{\mu}_{4,k} = \Sigma_{j=1}^{J_k} (Y_j^{(k)} - \bar{Y}^{(k)})^4 / J_k, \text{ and}$$

$$\hat{\delta}_{4,k} = \Sigma_{i=1}^{M} p_i^{(k)} (1-p_i^{(k)})\, (1-2p_i^{(k)})^2.$$

Step 5: Repeat steps 3 and 4 for items in AT2 and compute the statistic T_B (B for bias correction) according to the Equation 1.3.

Step 6: Perform the test for unidimensionality.

Stout's test for unidimensionality is given by

$$(1.4) \qquad T = (T_L - T_B)/\sqrt{2}.$$

Reject H_0: d = 1 if $T \geq Z_\alpha$, where Z_α is the upper 100(1−α) percentile for items in this case, subgroups tended to have examinees varying highly in the ability being tested. This lead to badly violating the assumption of local independence within subgroups. Hence the procedure performed badly in the d=1 case.

Correction for Increased Bias

It was observed that the subgroups of examinees would be more desirable if they were placed into subgroups on the basis of their scores on items that are not all difficult. This can be achived in the following way.

1. After selecting items into AT1, they are checked statistically to see if they are too easy as a group.

2. If so, they are replaced with items of highest loadings of opposite sign so that they are still as unidimensional as possible when d=1.

3.Otherwise, items are retained.

An Algorithm for Determining the size of Assessment Subtests

Prior to developing this algorithm, the size of the assessment subtests (M) was specified by the user prior to applying Stout's procedure for assessing dimensionality.

The proposed algorithm mechanically determines the size of assessment subtests AT1 and AT2 based on the magnitude of item loadings on the second factor.

Monte Carlo Simulation Studies

A large scale simulation study was conducted in both d=1 and d=2 cases. The purpose was to establish that Stout's procedure after undergoing correction for increased bias, and using the new algorithm to determine the size of AT1 and AT2, provides strong empirical evidence of the utility of the statistical test in a variety of test settings. More precicely two issues were addressed:

(a) how well the nominal level of significance specified by the user (α = .05) is approximated by the observed level of significance, and

(b) how well the power of the statistical test is maintained while undergoing the above-mentioned changes.

the standard normal distribution, α being the desired level of significance.

Note: In all the simulation studies, a part of the sample is used to perform factor analysis in order to place items into subtests AT1, AT2, and PT in Step 1, and the rest of the sample is used to compute the statistic in steps 2 through 6.

A Limitation of Stout's Procedure

The items of SAT-verbal test (confirmed as unidimensional) were divided into two groups. One group with items having a-parameter greater than 1.0 and the other group witn a-parameter less than 1.0. Stout's procedure was applied to assess dimensionality of each subgroup as if it was a test in itself. The results were markedly different.

Table 1

Rejection Rates per 100 Trials for d = 1 3PL Simulation Study Using Estimated Item Parameters of SAT Verbal Test With $\alpha = 0.05$

Discrimination	Number of	Number of examinees		
parameter	items	750	1000	2000
low a's ($0 \leq a_i \leq 1.0$)	41	4	0	3
high a's ($1.1 \leq a_i \leq 2.0$)	39	28	46	58

Reason for Increased bias

In Monte Carlo simulations studies, factor analysis was used as a data analytic tool to select items into AT1. Using principal axis factor analysis, items with high loadings of the same sign (either positive or negative) on the second extracted factor are selected into AT1. In the case of high a's, most often, very easy items tended to have highest loadings in magnitude. Consequently, the easiest items were put into AT1. Stout's procedure then, in an attempt to control for statistical bias due to short test lengths, puts the remaining easy items into AT2. Therefore, PT was left with difficult items remaining. Because examinees are grouped according to their score on PT, consisting of difficult or very difficult

The following results illustrate that using the proposed bias correction method together with the new algorithm to determine the size of assessment subtests has completely eliminated the excess bias due to high a's in Stout's statistic and improved the performance of Stout's procedure.

Table 2

Rejection Rates per 100 Trials for d = 1 3PL Simulation Study Using Estimated Item Parameters of Respective Tests With $\alpha = 0.05$

J	TESTS						
	SAT V	SAT V high a'S	ACT M	ACT E	ASVAB AS	ASVAB AR	ASVAB GS
	50	*39*	*40*	*50*	*25*	*30*	*25*
750	0	0	1	1	1	0	2
2000	0	0	1	2	1	0	13

Notes: Numbers in bold face represent the number of items used in the simulation study of the respective test.

J denotes the number of examinees simulated.

SAT V denotes the Scholastic Aptitude Test for verbal.

SAT V high a's denotes the Scholastic Aptitude Test where items have high discrimination parameter, namely $1.1 \leq a_i \leq 2.0$.

ACT M denotes the American College Test for mathematics usage.

ACT E denotes the American College Test for english usage.

ASVAB AS denotes the Armed Services Vocational Aptitude Battery for Auto Shop information.

ASVAB AR denotes the Armed Services Vocational Aptitude Battery for Arithmetic Reasoning.

ASVAB GS denotes the Armed Services Vocational Aptitude Battery for General Science.

Table 3

Rejection Rates per 100 Trials for d = 2, c = 0.2 3PL Simulation Study With $\alpha = 0.05$

	TESTS*					
Parameters	SAT V	ACT M	ACT E	ASVAB AS	ASVAB AR	ASVAB GS
	50	**40**	**50**	**25**	**30**	**25**
$N_1:N_2:N_3$	17:17:16	13:13:14	17:17:16	8:8:9	10:10:10	8:8:9
J	750 2000	750 2000	750 2000	750 2000	750 2000	750 2000
$\rho = 0.5$	94 100	87 98	68 91	65 97	84 98	68 100
$\rho = 0.7$	36 69	44 77	19 31	35 70	43 74	31 74

Notes: Numbers in bold face represent the number of items used in the simulation study of the respective test.

* the two artificial dimensions have item parameter distribution as that of the respective real test.

N_1 denotes the number of pure items of ability 1.

N_2 denotes the number of pure items of ability 2.

N_3 denotes the number of mixed items requiring the knowledge of both ability 1 and ability 2.

ρ denotes the correlation between the abilities.

J denotes the number of examinees.

Advantages of Stout's Procedure.

1. Unlike most other procedures, Stout's statistic and the procedure grew directly out of a meaningful conceptual definition of test dimensionality. That is, Stout's statistic T is designed to be sensitive only to dominant dimensions and not sensitive to item idiosyncracies.

2. The procedure is supported by an asymptotic theory.

3. It is nonparametric, thus avoiding the issues of parametric model correctness.

4. A major advantage of the procedure from the practitioner's point of view is that the computational requirements are modest and hence cost effective. For example, it takes 7 seconds to assses the unidimensionality of a 30 item test and 20 seconds for a 50 item test on CYBER 175.

5. Lastly, extensive Monte Carlo simulations for a wide variety of test lengths, and sample sizes, as also can be seen in Stout (Psychometrika, Dec 87) strongly support the validity of the procedure.