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

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 \( \theta \) usually assumed continuous and unidimensional random variable.

Assumptions Underlying the Item Response Theory Models:

1. Unidimensionality, \( d=1 \)
2. \( p(\theta) \) increases in \( \theta \), "monotonicity"
3. \( P(U_i = u_j | \theta) = \frac{1}{N} \sum_{i=1}^{N} P(U_i = u_j | \theta) \) for all \( u \) and \( \theta \)

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 \( \theta \) such that for all values of \( \theta \),

\[ \frac{1}{N(N-1)} \sum_{1 \leq i < j \leq N} |\text{cov}(U_i, U_j | \theta)| \neq 0. \]
This is to say that on the average the conditional covariances should be small. Stout further defines an empirical notion of unidimensionality consistent with Equation 1.1 as follows:

A test \((U_1, U_2, ..., U_N)\) is unidimensional if for all subtests \(((U_{k_1}, U_{k_2}, ..., U_{k_M})_k)\) of length \(M (< N)\) and all values of \(Y_p\),

\[
S_{M,N} = \frac{1}{M(M-1)} \sum_{1 \leq i < j \leq M} \text{cov}(U_{i,k}, U_{j,k} | Y) \neq 0
\]

where \(Y_p\) is the proportion correct on the long subtest complementary to \((U_{k_1}, U_{k_2}, ..., U_{k_M})\) 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.

Stout's Test for Assessing Unidimensionality

\[H_0: d = 1 \text{ vs } H_1: d > 1\]

The main underlying assumptions

1. Examinees are randomly selected from a large population.
2. Examinees respond to items independently 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.

Test Procedure

Step 1. Split \(N\) test items into 3 subtests: two short assessment subtests of length \(M\) each, and a long partitioning subtest of length \(n = N - 2M\).

\(\text{AT1}—\)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 purpose of AT1 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.

$$
\hat{\sigma}_{k}^{2} = \frac{\sum_{j=1}^{J_k} (Y_{(k)}^{(j)} - \bar{Y}_{(k)})^2}{J_k},
$$

is the usual variance estimate for subgroup $k$, where $Y_{(k)}^{(j)} = \frac{\sum_{i=1}^{M} U_{ijk}}{M}$, and $\bar{Y}_{(k)} = \frac{\sum_{j=1}^{J_k} Y_{(k)}^{(j)}}{J_k}$, and

$$
\hat{\sigma}_{U_{(k)}}^{2} = \frac{\sum_{i=1}^{M} (1-p_{i})^{2}}{M},
$$
is the unidimensional variance for subgroup \( k \), where \( p^{(k)} = \frac{\sum_{j=1}^{J} U_{ijk}}{J_k} \).

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 \( \hat{\sigma}_k^2 \approx \hat{\sigma}_{U,k}^2 \). Under \( H_1 \), however, the assumption of local independence will be badly violated leading to \( \hat{\sigma}_k^2 >> \hat{\sigma}_{U,k}^2 \).

Step 4: Compute the statistic \( T_L \) (L for long test) for items in AT1 subtest:

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

where \( S_k \) is the appropriate normalizing constant given by:

\[
S_k^2 = \left[ \hat{\mu}_{4,k}^4 - \frac{1}{4} \hat{\mu}_{4,k}^{(4,k)} \right] + \hat{\delta}_{4,k}^{(2,k)} M^2 + 2 \hat{\mu}_{4,k} \hat{\delta}_{4,k}^{(2,k)} M^2 \left[ \frac{1}{J_k} \right]
\]

where

\[
\hat{\mu}_{4,k} = \frac{\sum_{j=1}^{J} (Y_j^{(k)} - \bar{Y}^{(k)})^4}{J_k}, \text{ and}
\]

\[
\hat{\delta}_{4,k} = \frac{\sum_{i=1}^{M} p_i^{(k)} (1-p_i^{(k)}) (1-2p_i^{(k)})}{J_k}.
\]

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

\[
T = \frac{T_L - T_B}{\sqrt{2}}.
\]

Reject \( H_0: \ d = 1 \) if \( T \geq Z_{\alpha} \), where \( Z_{\alpha} \) is the upper 100(1-\( \alpha \)) percentile for \( \alpha \).
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 achieved 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 precisely two
issues were addressed:

(a) how well the nominal level of significance specified by the user ($\alpha =
.05$) 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, $\alpha$ 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 with 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 $\alpha = 0.05$

<table>
<thead>
<tr>
<th>Discrimination parameter</th>
<th>Number of items</th>
<th>Number of examinees</th>
</tr>
</thead>
<tbody>
<tr>
<td>low a's ($0 \leq a_i \leq 1.0$)</td>
<td>41</td>
<td>750</td>
</tr>
<tr>
<td>high a's ($1.1 \leq a_i \leq 2.0$)</td>
<td>39</td>
<td>28</td>
</tr>
</tbody>
</table>

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 $q = 0.05$

<table>
<thead>
<tr>
<th>J</th>
<th>SAT V</th>
<th>SAT V</th>
<th>ACT M</th>
<th>ACT E</th>
<th>ASVAB AS</th>
<th>ASVAB AR</th>
<th>ASVAB GS</th>
</tr>
</thead>
<tbody>
<tr>
<td>750</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>2000</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>13</td>
</tr>
</tbody>
</table>

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 $1.1 < a_i < 2.0$.

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>
<thead>
<tr>
<th>Parameters</th>
<th>SAT V</th>
<th>ACT M</th>
<th>ACT E</th>
<th>ASVAB AS</th>
<th>ASVAB AR</th>
<th>ASVAB GS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>50</td>
<td>40</td>
<td>50</td>
<td>25</td>
<td>30</td>
<td>25</td>
</tr>
<tr>
<td>$p = 0.5$</td>
<td>94</td>
<td>100</td>
<td>87</td>
<td>98</td>
<td>68</td>
<td>91</td>
</tr>
<tr>
<td>$p = 0.7$</td>
<td>36</td>
<td>69</td>
<td>44</td>
<td>77</td>
<td>19</td>
<td>31</td>
</tr>
</tbody>
</table>

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.

$p$ 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 idiosyncrasies.

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 assess 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.