ED 358 118 TM 019 887

AUTHOR Ang, Cheng; Miller, M. David

TITLE An Investigation of the Power of Stout's Test of Essential Unidimensionality.

PUB DATE Apr 93

NOTE 31p.; Paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993).

PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS Ability; Comparative Testing; Correlation; Item Response Theory; Mathematical Models; *Sample Size; Test Construction; *Test Length

IDENTIFIERS Null Hypothe is; Power (Statistics); *Stouts Procedure; *Unidimensionality (Tests)

ABSTRACT

The power of the procedure of W. Stout to detect deviations from essential unidimensionality in two-dimensional data was investigated for minor, moderate, and large deviations from unidimensionality using criteria for deviations from unidimensionality based on prior research. Test lengths of 20 and 40 items and sample sizes of 700 and 1,500 were studied. The power of Stout's procedure was directly related to the deviation from unidimensionality based on deviation areas. Deviation areas were inversely related to the correlation between the dominant ability and the reference composite. When the sample size or test length increased, the power of Stout's procedure also increased. In general, Stout's procedure had sufficient power to reject the null hypothesis of essential unidimensionality if 10 to 20 percent of the items were dimensionally distinct from the rest of the items. Results indicate that for minor deviation from unidimensionality, the rejection rates of Stout's procedure were not near the nominal level of 5 percent. For moderate and large deviations from unidimensionality, Stout's procedure had power to reject the null hypothesis of essential unidimensionality, especially if the sample size was 1,500 and the test length was 40. Twelve figures and 10 tables illustrate the discussion. (SLD)

AN INVESTIGATION OF THE POWER OF STOUT'S TEST OF ESSENTIAL UNIDIMENSIONALITY

CHENG ANG
M. DAVID. MILLER

ABSTRACT

The power of Stout's procedure to detect deviations from essential unidimensionality in two-dimensional data was investigated for minor, moderate, and large deviations from unidimensionality. The criteria used in the categorization of deviations from unidimensionality were based on Shepard, Camilli, and Williams's categorization of area measures of item bias.

The power of Stout's procedure was directly related to the deviation from unidimensionality based on deviation areas. Deviation areas were inversely related to the correlation between the dominant ability and the reference composite. When the sample size increased, the power of Stout's procedure also increased. The power for 40-item tests was higher than for 20-item tests. When the proportion of items loaded on the minor dimension was 20%, the power was the highest. Although the power for the 20% condition was higher than the 100% condition, the correlations $\rho_{\gamma,\theta 1}$ for the 20% condition were also extremely high. For the 10% and 20% conditions, even when the correlations $\rho_{\gamma,\theta 1}$ were near 1.00, the rejection rates were high.

In general, Stout's procedure had sufficient power to reject the null hypothesis of essential unidimensionality if 10% to 20% of the items were dimensionally distinct from the rest of the items. This is because only 10% to 20% of the items are being selected into the subtest (AT1) used in testing essential unidimensionality. When AT1 is dimensionally distinct from the rest of the items, Stout's null hypothesis of essential unidimensionality will be rejected.

The results of this study indicate that for minor deviation from unidimensionality, the rejection rates of Stout's procedure were not near the nominal level of 5%. For moderate and large deviation from unidimensionality, Stout's procedure had power to reject the null hypothesis of essential unidimensionality, especially if the sample size was 1,500 and the test length was 40. Further studies are recommended.

AN INVESTIGATION OF THE POWER OF STOUT'S TEST OF ESSENTIAL UNIDIMENSIONALITY

Introduction

Stout's procedure and the concept of essential unidimensionality have been described in detail (Nandakumar, 1991; Stout, 1987, 1990). Although the power of Stout's procedure has been studied (Nandakumar, 1991; Stout, 1987), all conditions manipulated. were not conducted with known minor, moderate, and large deviations from unidimensionality. Also, the effect of the proportion of items loaded on the minor dimension and the effects of test lengths have not been systematically studied.

Purpose

The purpose of this study was to investigate the power of Stout's procedure to detect deviations from essential unidimensionality in two-dimensional data. The specific questions were as follows:

1. How do minor, moderate, and large deviations from unidimensionality affect the power of Stout's procedure for testing essential unidimensionality?

2. How does the proportion of items loaded on the minor dimension affect the power of Stout's procedure for testing essential unidimensionality?

3. How does test length affect the power of Stout's procedure for testing essential unidimensionality?

4. How does sample size affect the power of Stout's procedure for testing essential unidimensionality?

Design of the Study

Test Length and Sample Size

In the present study, the test length and sample size each have two levels. The test lengths studied were 20 (a short test) and 40 (an average-length test). Small and large sample sizes of 700 and 1,500 were studied. Hattie (1984) stated that sample sizes smaller than 300 tended to be unstable for latent trait procedures. In addition, large sample sizes (>5,000) might result in inappropriate rejection rates.

Item-Parameters

The item test parameters of the two-dimensional model with one major and one minor dimension were used in generating item response data. The first dimension was the major dimension that the test was purported to measure and the second dimension was the minor dimension. The influence of the second dimension on each item was relatively weak compared to that of the first dimension. The means and variances of a_1 and a_2 reflected the degree to which their respective traits influenced item scores. An item with a large a_1 and a small a_2 was much more heavily influenced by θ_1 than by θ_2 and vice-versa (Nandakumar, 1991). Both θ_1 and θ_2 were normal with mean zero and variance equal to one. The correlation between the abilities was set at zero.

A preliminary investigation using Nandakumar's ξ (1991) to control the weight of the major dimension relative to the weight of the minor dimension was carried out. The item test parameters a_1 and a_2 were computed by varying μ, σ and ξ in the following expressions:

$$a_1 \sim N((1-\xi)\mu, (1-\xi)^{1/2}\sigma)$$

$$a_2 \sim N(\xi\mu, \xi^{1/2}\sigma)$$

$$a_1 + a_2 \sim N(\mu, \sigma) \qquad (1)$$

With $\mu = 1.07$ and $\sigma^2 = 0.16$, Stout's test of essential unidimensionality had little power. Even when the weight of the minor dimension was the same as the major dimension ($\xi = 0.5$), the rejection rates were less than 20%. With $\sigma^2 = 0.64$ however, there was substantial power even at $\xi = 0.3$. Using Nandakumar's ξ requires a large σ^2 because increasing ξ with a constant μ and σ^2 led to a decrease in $\sigma_{a1} = (1- \xi)^{1/2}\sigma$ and $\mu_{a1} = (1-\xi)\mu$. Unless σ_{a1}^2 is relatively large, the mininum of the variance of a_1 and a_2 will also be small (β) and lead to little power (Nandakumar, 1991). To avoid using a large σ^2 and hence a large range, in this study a small σ^2 was studied and the effect of the reduction in σ_{a1} and μ_{a1} (due to ξ) was controlled by holding both σ_{a1} and μ_{a1} constant.

In this study, the values of a_1 were fixed across conditions; that is, only one set of a_1 was used across deviation areas for each test length. For the 40-item tests, the a_1 parameters used in this study were the discrimination parameters of a 40-item ACT math test reported by Drasgow (1987). The mean and sigma of a_1 were 1.09 and 0.35, and a_1 ranged from .40 to 2.00. For the 20-item test, a_1 parameters were selected from the 40-item test parameters with mean 1.09, sigma 0.36 and a_1 ranged from 0.40 to 2.00. The mean and sigma for a_2 were W(1.09) and $W^{1/2}$(0.35), where W is the weighting factor similar to Nandakumar's (1991) use of ξ (ξ = (W / 1 + W)); that is, the μ and σ of a_2 were weighted by $W\mu$ and $W^{1/2}\sigma$ of a_1 instead of $\xi\mu$ and $\xi^{1/2}\sigma$ of the common a_s in Nandakumar (1991). Although W and ξ are basically the same, the purpose of using W was to keep a_1 the same across deviation areas. Because a_1 did not change, a_1 and a_2 were equal when the weight W = 1.00 (as opposed to Nandakumar's ξ = 0.5).

To compute the parameters for a_2, the parameters for a_1 were randomly rearranged using random numbers for both the 40- and 20-item tests. The purpose of rearranging the values of a_1 was to use the new a_1 for computing the weighted a_2 so that a_2 would be statistically independent of the original a_1 (the original a_1 was used for the major dimension). The item parameters of a_2 were computed by varying W on the new a_1 (e.g., a_2 = 0.34 * new a_1). W ranged from .34 to 0.90 to explore the desired deviation areas (this will be described further under the deviation areas section). Because only one set of random numbers was used for each test length to generate the new a_1, the item parameter a_2 for each item will have the same value across deviation areas if multiplied by 1/W.

The difficulty parameters reported by Drasgow (1987) for the ACT math test were also used in this study. The values reported by Drasgow were used for both b_1 and b_2 (b_1 has the same value as b_2). Because Drasgow (1987) only reported item difficulties for a 40-item test, b_1 and b_2 for the 20-item tests were selected from parameters for the 40-item test. The mean and standard deviation for b_1 and b_2 were about the same for the 20-item and the 40-item test parameters: μ of b_1 and b_2 were 0.50 and σ of b_1 and b_2 were 0.61 for both test lengths. The range, however, differed: for the 40-item test parameters, the range was from -1.02 to 1.50; for the 20-item test, the range was from -.60 to 1.50. For each test length, the same values of b_1 and b_2 were used across deviation areas to ensure that variation in deviation areas was not confounded with fluctuations in the difficulty parameters.

Proportion of Items

The proportion of items loaded on the minor dimensions had three levels: 10%, 20% and 100%.

The 10% and 20% levels were studied because many achievement tests have 5% to 29% of the items loaded on the second dimension (Ackerman, 1987). For example, a math test might consist of 10% word problems, thus requiring the ability to comprehend sentence structure. The 100% level was studied because it is common to have all test items contaminated by a second trait, although the contamination might be relatively weak compared to the influence of the first dimension (Nandakumar, 1991). For example, the ability of an examinee to answer all the math test items might be influenced by the examinee's ability to understand the instructions in English.

For the three proportions of items loaded on the second dimension, the parameters (a_1 and b_1) for the major dimension were the same across deviation areas for each test length. For the minor dimension, when 10% and 20% of the items loaded on the minor dimension, those items had the same a_2 and b_2 as some matched items in the 100% condition; that is, 10% and 20% of the items were selected from the 100% condition and the rest of the item loadings on the minor dimensions were set to 0.00. The selection of items for the 10% and 20% of the items loaded on the minor dimension will be discussed further under the deviation areas section.

Analytical Estimates

Equations for estimating the unidimensional item test parameters of the two-parameter model from the trait and item test parameters of a two-dimensional compensatory model have been established by Wang (1986). Because the data in this study were generated based on a bivariate extension of the 2PL model with compensatory abilities (equation 3.5) and the dimensions were assumed to be uncorrelated, Wang's (1986) special case formula was used in the estimation of the parameters of the unidimensional two-parameter model:

$$\hat{a}_j = \frac{W_1' a_j}{\sqrt{(1 + a_j' W_2 \; W_2' a_j)}} \qquad (1)$$

and

$$\hat{b}_j = \frac{b_j\sqrt{a_j'a_j}}{W_1'a_j} \qquad (2)$$

where

w_1(p x 1) is the first eigenvector of the matrix A'A;
w_2(p x 1) is the second eigenvector of the matrix A'A;
A(n x p) is the matrix of p discrimination parameters for each of n items;
a_j(1 x p) is the ith row of A, a vector of discriminations for item j; and
b_j(1 x p) is the ith row of b, a vector of difficulties for item j.

Categorization of Deviations

After the parameters of the unidimensional model were analytically estimated using Wang's procedure, differences between the analytical estimations and the first dimension of the true parameter values were computed using the unsigned area (UA) between the two item characteristic curves (ICCs) (Raju, 1988). The UA was computed by using

$$UA = \left| \frac{2(\hat{a} - a_1)}{Da_1\hat{a}} \ln(1 + \exp\left(\frac{Da_1\hat{a}(\hat{b} - b_1)}{(\hat{a} - a_1)}\right) - (\hat{b} - b_1) \right| \qquad (3)$$

where

a_1 is the discrimination parameter for the major dimension,
a is the discrimination parameter for the estimated dimension,
b_1 is the difficulty parameter for the major dimension, and
b is the difficulty parameter for the estimated dimension.

The area was then averaged over all the items loaded with two dimensions for each test. The deviations were grouped into three categories based on the average area: minor, moderate, and large deviations.

The criteria used to determine the three levels of deviations were based on the criteria used by Shepard, Camilli, and Williams (1985) in the categorization of bias between two groups. Their criteria were based on the differences between the difficulty parameters, b_1-b_2, of the two groups. When b1-b2 was less than .20, an item was classified as unbiased; when b1-b2 was between .20 and .35, an item was classified as weakly biased; and when b1-b2 was greater than .35, an item was classified as moderately biased. Their rationale for the categorization of biases was based on the examination of actual data (Shepard et al., 1985).

Since Raju's (1988) area procedure for the Rasch model between two ICCs was UA = $|b_1$-$b_2|$, the absolute value of the index used by Shepard et al. (1985) would be equivalent to that of Raju's area. Thus, in this study, when the area was less than .20, it was classified as a minor deviation; an area between .20 and .35 was classified as a moderate deviation; and an area greater than .35 was classified as a large deviation. Table 1 shows the characterizations of the three categories of deviations frcm unidimensionality.

Table 1

Three Deviations from Unidimensionality and Criteria for Cut-Off

Raju's UA	Deviation Category
< .20	Minor
≥ .20 ≤ .35	Moderate
> .35	Large

From the three categories of deviations from unidimensionality, six unique areas were chosen for the generation of data and testing of the hypothesis of essential unidimensionality. The areas chosen were 0.19 for a minor deviation; 0.28, 0.31 and 0.34 for a moderate deviation; 0.37 and 0.40 for a large deviation. The area of 0.19 represented the maximum area for a minor deviation area. The area of 0.28 was at the median (approximately) of the moderate deviation area. The rest of the deviation areas were an increment of 0.03 from 0.28 through the large deviation area of 0.40.

Deviation Areas

In this study, μ and σ of the difficulty parameters (b), and μ and σ of a_1 were fixed, therefore, the variation in deviation areas was determined by the size of a_2 relative to a_1 as controlled by the weighting factor W. As W increased, a_2 and the deviation areas also increased. Because deviation areas 0.19, 0.28, 0.31, 0.34, 0.37 and 0.40 were fixed apriori, W was explored from a range of 0.34 to 0.90 to create the six deviation areas; that is, different values of W were used until a pre-specified deviation area was obtained.

For each test length, all six deviation areas had the same a_1 parameters, but a_2 parameters were weighted by W.

To ensure the same deviation areas across test lengths, some minor changes were made in the a_1 parameters of the 20-item test. When the a and b parameters of the 20-item test were selected from the 40-item test (with the same mean, variance and W as the 40-item test), the average deviation areas for the 20-item test were slightly lower than the 40-item test when the deviation area was 0.40 (eg., instead of 0.40 in the 40-item test, it was 0.39). Therefore, minor changes were made in the values of a_1 (e.g, instead of the value of a_1 for item 13 = 0.62, it was changed to 0.66). The changes were made only for the large deviation area of 0.40 (W = 0.90 as in the 40-item test) and any decrease for an item was compensated for by the same increase to another item or vice-versa (to ensure the same mean and variance). Once the changes had been made and the 20-item test had the same W, same deviation area (0.40), and about the same mean, variance, and range for a_1 as in the 40-item test, the rest of the deviation areas for the 20-item tests were weighted by the same W as the rest of the deviation areas for the 40-item tests; that is, given the same W, all the deviation areas for the 20-item tests were the same as the 40-item tests with up to 0.01 rounding errors.

To ensure the same deviation areas across the proportion of items loaded on the minor dimension, the item test parameters for the major dimension of the 10% and 20% conditions were the same as the 100% condition for each test length. For the minor dimension of the 10% and 20% condition, the same proportion of items was selected from the minor dimension of the 100% condition. Those items not loaded on the minor dimension were set to zero and only those items loaded on the minor dimension we e computed for the deviation areas. Items loaded on the minor dimension of the 10% and 20% conditions were selected only from the deviation area of 0.40 (of the 100% condition). After the items that averaged to about 0.40 deviation areas had been selected (a few minor adjustments were made on the a_1 item-parameters to ensure the same deviation areas, especially for the 10% conditions), the same items were used for computing the other deviation areas, and the second dimension of the other deviation areas was weighted by the same W as used in the 100% condition. Therefore, for all deviation areas, the 10%

and 20% conditions would have the same weight (W) as the 100% condition. For 10% and 20% of the items loaded on the minor dimension, the deviation areas of those items loaded were about the same as the 100% condition with rounding errors of less than 0.01.

Given a fixed deviation area, W was the same across test lengths and the proportion of items loaded on the minor dimension. In this study, the final levels of W resulting in the six deviation areas are shown in Table 2. For the upper limit of the minor deviation area, a_2 was about one third the size of a_1, and for the upper limit of the moderate deviation area, a_2 was about two-thirds the size of a_1.

Table 2

Level of W and the Six Deviation Areas

Deviation Area	0.19	0.28	0.31	0.34	0.37	0.40
W	0.34	0.54	0.62	0.70	0.80	0.90

Item Response Data Generation

The item-parameters for the two test lengths and the three proportions of items loaded on the minor dimension were used to generate item response data. The same item-parameters were used for the 700 and 1,500 sample sizes. Both the θ_1 and θ_2 were generated from a normal distribution with mean zero and variance equal to one, and θ_1 and θ_2 were independent. The means and variances of θ_1 and θ_2 were the same across replications. Two levels of sample size, two levels of test length, and three proportions of items loaded on the second dimension were crossed with each other to create 12 unique conditions. Table 3 presents the design for the sample sizes, the test lengths, and the proportions of items loaded on the minor dimension. The generation of item responses was repeated 100 times for each of the six deviation areas, totaling 7200 data sets. Table 4 shows replication of the three categories of error based on fixed deviation areas for the 12 conditions.

Test of Hypothesis

For each item response data set generated, Stout's nonparametric procedure was used to test the hypothesis of H_o: d = 1 versus H_1: d > 1; that is, whether the data were essentially unidimensional. Stout's (1991) dimensionality testing program (DIMTEST) was used. In DIMTEST, AT1 items can be selected either by expert's opinion or by factor analysis of tetrachoric correlations. In this study, factor analysis was used and the sample size used in factor analysis was 700 and 1,500 (same sample size as DIMTEST). Because the purpose of factor analysis was to select items for AT1 in DIMTEST, 10 factor analyses were performed for each unique condition (using different replications) and the factor analysis that produced the most dimensionally distinct items (as determined from examining the item parameters) for AT1 was used ᶠor the rest of the replications; that is, the same AT1 items were used for 100 replications. Thus, differences across the replications could not be attributed to the use of different item parameters (selected for AT1) being used in Stout's test of essential unidimensionality.

Table 3

Design for Sample Size, Test Length, and Proportion of Items Loaded on the Second Dimension

Sample Size	Test Length	Prop. of Items	Condition
700	20	10%	1
		20%	2
		100%	3
	40	10%	4
		20%	5
		100%	6
1,500	20	10%	7
		20%	8
		100%	9
	40	10%	10
		20%	11
		100%	12

Table 4

Replication of the Three Categories of Error for the 12 Conditions

		Conditions
Category:	Area	1 2 3................12
Minor	0.19	100 100 100
Moderate	0.28	100 100 100
	0.31	100 100 100
	0.34	100 100 100
Large	0.37	100 100 100
	0.40	100 100 100

Simulation Models

Both the univariate 2PL model and the bivariate extension of the 2PL model with compensatory abilities were used in the generation of data. Two dimensional items were generated using the following equation:

$$P_i(\theta_1,\theta_2) = \frac{1}{1+\exp[-1.7[a_{1i}(\theta_1-b_{1i}) + a_{2i}(\theta_2 - b_{2i})]]} \qquad (5)$$

where

θ_1 and θ_2 are the ability parameters for dimensions one and two,
a_{1i} and a_{2i} are the discrimination parameters for item i on the two dimensions, and
b_{1i} and b_{2i} are the difficulty parameters for item i.

Nandakumar (1991) showed that when a_2 and b_2 of the second dimension are zero, equation 3.5 reduces to a unidimensional 2PL logistic model with respect to θ_1. Therefore, when 10% and 20% of the items were two-dimensional, unidimensional items were simulated by using the unidimensional 2PL model

$$P_i(\theta_1) = \frac{1}{1+\exp(-1.7a_i(\theta_1-b_i))} \qquad (5)$$

Results

Table 5 shows the distribution of rejection rates across all conditions.

Table 5

The Distribution of Rejection Rates Across All Conditions.

		Sample Size			
		700		1,500	
Areas	Proportion	20*	40*	20*	40*
0.19	100%	19	11	34	36
	20%	19	30	25	52
	10%	4	13	24	32
0.28	100%	26	28	43	68
	20%	64	90	91	98
	10%	19	57	29	88
0.31	100%	24	41	43	82
	20%	66	92	99	98
	10%	15	75	29	94
0.34	100%	35	59	58	99
	20%	87	99	99	100
	10%	19	87	44	98
0.37	100%	34	80	63	100
	20%	97	100	100	100
	10%	27	99	60	100
0.40	100%	53	88	76	100
	20%	100	100	100	100
	10%	27	100	62	100

Note. * refers to test length.

Deviation Areas

As shown in Table 6, the average rejection rate for the minor deviation was about 25%. Table 5 shows that regardless of sample size, proportion of items loaded on the second dimension, and test length, none of the rejection rates for minor deviations were above 52% and the lowest rejection rate was 4%.

Table 6

Mean Rejection Rates for Each Deviation Area

Deviation Area					
0.40	0.37	0.34	0.31	0.28	0.19
83.83	80.00	73.67	63.17	58.42	24.92

For the moderate deviations, the rejection rates averaged about 65% (Table 6). Table 5 shows that the rejection rates for the moderate deviations ranged from 19% to 100%. For the large deviations, the average rejection rate was 82%, and the rejection rates ranged from 27% to 100%.

Effect of Test Length

As shown in Table 7, 20-item tests averaged a 50% rejection rate and 40-item tests averaged a 78% rejection rate.

Table 7

Mean Rejection Rates for Each Test Length

Test Length	
20	40
50.39	77.61

Proportion of Items

As shown in Table 8, the 10% and 100% conditions averaged about 54% rejection rates, but the 20% condition averaged about 84%.

Table 8

Mean Rejection Rates for Each Proportion of Items Based on Duncan's Multiple Range Test

Proportion of Items		
100%	20%	10%
54.250	83.583	54.167

Sample Size

Tnble 9 shows that the 700 subject condition averaged about a 55% rejection rate and the 1,500 subject condition averaged about a 73% rejection rate.

Table 9

Mean Rejection Rates for Each Sample Size

Sample Size	
700	1,500
55.11	72.89

The Power of Stout's Procedure

The power curves from Figures 4.1 through 4.12 show that for all test lengths, sample sizes, and proportions of items loaded on the minor dimension, the power increased as the deviation area increased.

In general, the rejection rates for each condition were directly related to the size of the deviation areas.

Discussion

The results of this study are discussed in relation to the correlation between the ability of the major dimension (θ_1) and the reference composite ability (γ) investigated in the preliminary investigation. In this investigation, the population correlation between γ and θ_1 was computed for each combination of deviation area, test length, and proportion of items on the second dimension. The population correlation between the reference composite (γ) and θ_1

$$\rho_{\gamma,\theta_1} = \frac{W_1}{(W_1^2 + W_2^2)^{0.5}} \qquad (6)$$

where
w_1(p x 1) is the first eigenvector of the matrix A'A, and
w_2(p x 1) is the second eigenvector of the matrix A'A

was derived from the estimated reference composite of Wang (1986),

$$\hat{\gamma}_j = \theta_j W_1 \qquad (7)$$

where
θ_j(1 x p) is the jth row of the matrix θ, and
w_1(p x 1) is the first eigenvector of the matrix A'A.

A low correlation means that the data were heavily influenced by the minor ability and the ability of interest (the major dimension) did not correspond to the reference composite. A high correlation means the influence of the minor ability was very mild and the ability of interest (major dimension) was consistent with the reference composite.

Proportion of Items

When the proportion of items loaded on the second dimension was 100%, the results of a preliminary investigation in Table 10 showed that the correlations between θ_1 of the simulated data and the reference composite were inversely related to the size of the deviation areas; that is, with both test lengths, high deviation areas yielded lower correlations and low deviation areas yielded higher correlations. The rejection rates were also directly related to the deviation areas. Looking at each of the deviation areas for 100% of the items loaded on the second dimension and when the correlation was high, such as 0.954, Stout's procedure rejected on the average of 27% for 20-item tests and 24% for 40-item tests. When the correlation was 0.75, the rejection rate of Stout's procedure was 65% on the average for 20-item tests and 94% for 40-item tests. Although there was no prior criterion for $\rho_{\gamma,\theta 1}$ for defining essential unidimensionality and non-essential unidimensionality, a $\rho_{\gamma,\theta 1}$ of 0.954 seems high and, thus, the data should be essentially unidimensional. As a consequence, the null hypothesis of essential unidimensionality should not be rejected. The relatively high rejection rates at a minor deviation area for both 20-item and 40-item tests show that Stout's procedure might have too much power in rejecting the null hypothesis of essential unidimensionality.

For the 20% of the items loaded on the second dimension condition, the results in Table 10 show that the correlations between theta 1 of the simulated data and the reference composite remained near the 0.99 or 0.98 levels, regardless of the deviation area or test length. Although the rejection rates were related to the deviation areas, the relationship between the correlation $\rho_{\gamma,\theta 1}$ and the deviation area was very mild. Because $\rho_{\gamma,\theta 1}$ was very high over all the deviation areas, Stout's null hypothesis of essential unidimensionality should not be rejected. In this study, however, when 20% of the items loaded on the minor trait, the rejection rates for minor, moderate, and large deviation areas were very high and some of the rejection rates were 100%. One reason for these high rejection rates was the selection of items into subtest AT1 through factor analysis. Factor analysis selects the M items into AT1 that load most heavily either positively or negatively on the second extracted factor (i.e., the selected items are dimensionally distinct from the rest of the items), resulting in the possible selection of most of the 20% ($M \leq N/4$) items that loaded on the second factor. Because the rest of the items were not loaded on the second factor, the selection of items for AT2 might not have had the same difficulty distribution as in AT1. Thus, examinees within each subgroup of PT were not likely to be approximately equal on the dominant trait measured by the test, which resulted in high rejection rates.

Table 10

Correlation Between Theta 1 and the Reference Composite

		Deviation Areas					
Test Length	% of Items	0.19	0.28	0.31	0.34	0.37	0.40
20	100	0.954 (27%)	0.891 (35%)	0.861 (34%)	0.829 (47%)	0.788 (49%)	0.747 (65%)
	20	0.998 (22%)	0.994 (78%)	0.992 (83%)	0.990 (93%)	0.986 (99%)	0.982 (100%)
	10	0.999 (14%)	0.999 (24%)	0.998 (22%)	0.997 (32%)	0.996 (44%)	0.995 (45%)
40	100	0.954 (24%)	0.891 (48%)	0.861 (62%)	0.829 (79%)	0.788 (90%)	0.747 (94%)
	20	0.997 (41%)	0.994 (94%)	0.991 (94%)	0.989 (99%)	0.985 (100%)	0.980 (100%)
	10	0.999 (23%)	0.998 (73%)	0.997 (85%)	0.996 (93%)	0.995 (100%)	0.994 (100%)

Note. The value in each of the parentheses is the average rates of the two sample sizes corresponding to the correlation $\rho_{\gamma,\theta 1}$.

Similar to the results obtained with the 20% condition; the correlations between theta 1 of the simulated data and the reference composite of the 10% condition was at the 0.99 level regardless of the deviation area or test length. The rejection rates for the 10% condition, however, were also relatively high. The high rejection rates for the 10% condition might be the result of the same factor as the high rejection rates for the 20% condition.

Although the rejection rates for the 20% condition were higher than for the conditions in which 100% of the items loaded on the second dimension, the correlation $\rho_{\gamma,\theta 1}$ for the 20% condition was much higher than for the 100% condition. This shows that the high rejection rates for the 20% condition (and for the 10% condition) might be the result of the weakness of Stout's procedure in selecting dimensionally distinct AT1 items in DIMTEST. Because the scores on AT1 are used to compute Stout's statistics, if AT1 items are dimensionally distinct from the rest of the items, the null hypothesis of Stout's procedure will be rejected. This can be a problem. If no dimensionally distinct items are present when all items load on both dimensions, Stout's null hypothesis may not be rejected. If there is a small proportion of dimensionally distinct items such as 10%, even when the weight on the second dimension is very weak, the null hypothesis may be rejected.

Test Length

Although the preliminary investigation showed that the two test lengths have about the same levels of correlation ρ_{γ,θ_1} over deviation areas, the finding of the present study suggests that Stout's procedure has more power in rejecting the null hypothesis of essential unidimensionality with longer tests than with shorter tests; that is, the power increased from the 20-item tests to the 40-item tests under moderate and large deviation areas, sample sizes, and proportions of items loaded on the second dimension. Although for minor deviation areas with 100% of the items loaded on the minor dimension, the 20-item test had slightly more power than the 40-item test; the result might be due to random error (the increase was very mild). In general, the result of this study is consistent with Nandakumar's observation (1991).

Sample Size

As shown in this study, when the sample size increased, the power of Stout's procedure also increased. The results were consistent with those of Stout (1987) and Nandakumar (1991). Larger sample sizes not only had an advantage over smaller sample sizes in terms of power using Stout's procedure, but larger sample sizes also provided a more stable estimation in factor analysis (Gorsuch, 1983) and thus, factor analysis selected better dimensionally distinct items for AT1 in DIMTEST.

Deviation Areas

In general, for all proportions of items loaded on the second dimension, test lengths, and sample sizes, as the deviation area increased, the rejection rate also increased; that is, the rejection rate for each condition is directly related to the deviation area. Although the rejection rates were directly related to deviation areas, μ, σ^2, and the range of a_s and b_s were fixed in this study. The 'mpact of μ, σ^2 and the range of a_s and b_s on deviation areas and the power of Stout's procedure need to be further explored with variations in these parameters.

The Power of Stout's Procedure

Although factor analysis is merely a data analytic technique for obtaining AT1 items that are as dimensionally distinct from the rest of the items as possible (Stout, 1987), a preliminary study found that when the deviation areas were minor and moderate, the power of Stout's procedure fluctuated as a function of the items selected by the factor analysis for AT1; that is, the more dimensionally distinct AT1 items tended to have higher rejection rates. Also, factor analysis did not select the most dimensionally distinct items for AT1 when the sample size used was small (500) and the test length was 40 items. To avoid the possibility that AT1 items selected by factor analysis might not be the most dimensionally distinct items, factor analysis was performed on the sample sizes of 700 and 1,500 (same as in DIMTEST) and attempts were made to ensure that items selected by factor analysis for AT1 were as dimensionally distinct from the rest of the items as possible; that is, 10 factor analyses were performed on the data sets for each condition and the factor analysis that yielded the most dimensionally distinct items for AT1 was used. In this study, given a fixed condition, the same AT1 items (the most dimensionally distinct set of items yielded by factor analysis) were used for 100 replications.

The results of this study showed that the power of Stout's procedure in rejecting the null hypothesis of essential unidimensionality was conditioned on sample size, test length, proportion of items loaded on the second dimension and deviation area. The power for each condition was directly related to the deviation area; that is, the larger the deviation area, the greater the power. A sample size of 1,500 had more power than a sample size of 700, and a 40-item test had more power than a 20-item test. In general, the power of Stout's procedure was relatively low for 20-item tests with 700 examinees but relatively high for 40-item tests with 1,500 examinees and this was true for all proportions of items loaded on the minor dimension.

Comparison to Previous Studies

Although Stout (1987) studied strictly unidimensional data and two dimensional data of equal weight (two equal dominant dimensions), this study only examined two-dimensional data with one major and one minor dimension. When the weight of the second dimension was large (large deviation areas), the power in this study was comparable to the results of Stout's (1987) two-dimensional data.

Nandakumar (1991) also examined the power of Stout's test of essential unidimensionality. But, the weight of the second dimension in this study was based on the weighting factor W, as opposed to ξ in Nandakumar (see Chapter 3); that is, the distributions of a_1 and b_1 of the major dimension were the same rega·dless of the weight of the second dimension. Because the major dimension was kept constant, there was no confounding of σ_{a1} and μ_{a1} across deviation areas. In contrast, Nandakumar (1991) generated data where σ_{a1} and μ_{a1} decreased as ξ increased. Because there was no reduction of σ_{a1} and μ_{a1} across deviation areas, the power in this study was much higher than the power in Nandakumar's study.. The item parameters in this study were fixed across conditions, thus the results are easier to interpret across sample size and the proportion of items loaded on the second dimension than in Nanadakumar (1991).

Limitation of the Present Study

Although two-dimensional data were used in the present study, the dimensions were assumed to be uncorrelated and guessing was not taken into account. Other correlations between dimensions, other μ, σ^2, other ranges of a_s and b_s, other sample sizes, other test lengths, and other proportions of items loaded on the second dimension might lead to different results.

In this study, many factor analyses were performed for each condition to ensure that items selected for AT1 were as dimensionally distinct from the rest of the items as possible. In practice with real data, this may not be possible because the data set may not be large enough to perform many factor analyses. Therefore, when working with real data, experts' opinions may be used in selecting data when appropriate. If factor analysis is used, care should be undertaken to ensure that the AT1 items are dimensionally distinct from the rest of the items by ensuring that only the highest a_1 with the lowest a_2, or vice versa, is chosen for AT1 items (Stout, personal communication).

Conclusion

The results of this study indicated that the power of Stout's procedure is directly related to the deviation areas. Deviation areas were inversely proportional to the correlation between the dominant ability and the reference composite. When the sample size increased, the power of Stout's procedure also increased. The power of Stout's procedure for 40-item tests was higher than for 20-item tests. When the proportion of items loaded on the minor dimension was 20%, the power was the highest. Although the power for the 20% condition was higher than for the 100% condition, the correlation $\rho_{\gamma,\theta 1}$ for the 20% condition was also extremely high. For the 20% and 10% conditions, even when the correlations $\rho_{\gamma,\theta 1}$ is near 1.00, the rejection rates can be high.

In general, Stout's (1987) procedure had sufficient power in rejecting the null hypothesis of essential unidimensionality if the combination of $\sigma_{a1}{}^2$ and μ_{a1}, and $\sigma_{a2}{}^2$ and μ_{a2} were such that about 10% to 20% of the items selected into AT1 (under all conditions) were dimensionally distinct from the rest of the items. If AT1 was dimensionally distinct from the rest of the items, then Stout's null hypothesis of essential unidimensionality would be rejected.

The results of this study indicate that for minor deviation areas, the rejection rates of Stout's procedure were not near the nominal level of 5%. For moderate and large deviation areas, Stout's (1987) procedure had sufficient power in rejecting the null hypothesis of essential unidimensionality, especially if the sample size was 1,500 and the test length was 40. The appropriateness of essential unidimensional data for unidimensional IRT estimation is unknown.

Further Research

The impact of μ, σ^2 and the range of a_s and b_s on deviation areas also need to be further explored. The results of this study and $\rho_{\gamma,\theta 1}$ imply that Stout's procedure may be too powerful; therefore, some adjustments in Stout's procedure need to be undertaken. A test for essentially unidimensional data could become a test for the appropriateness of unidimensional IRT estimation with two-dimensional data.

The appropriateness of essentially unidimensional data for equating and adaptive testing has not been explored. Other variables that may influence the power of Stout's procedure, such as the direction of items and guessing, may need to be systematically studied. Lastly, only one major and one minor ability were studied here. Preliminary investigation showed Stout's procedure had more power when more than one minor ability was present. Therefore, the power of Stout's procedure based on one major and many minor abilities may need further research.

REFERENCES

Ackerman, T. A. (1987, April). The use of unidimensional item parameter estimations of multidimensional items in adaptive testing. Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C.

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. Applied Psychological Measurement, 13, 113-127.

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

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. Multivariate Behavioral Research, 19, 49-78.

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

Nandakumar, R. (1987). Refinement of Stout's procedure for assessing latent trait unidimensionality (Doctoral dissertation, University of Illinois at Urbana-Champaign, 1987). Dissertation Abstracts International, 49, 01A.

Nandakumar, R. (1991). Traditional dimensionality vs. essential dimensionality. Journal of Educational Measurement, 28, 1-19.

Nandakumar, R., & Stout, W. F. (in press). Refinement of Stout's procedure for assessing latent trait unidimensionality. Journal of Educational Statistics.

Raju, N. (1988). The area between two item characteristic curves. Psychometrika, 53, 495-502.

Shepard, L., Camilli, G., & Williams, D. (1985). Validity of approximation techniques for detecting item bias. Journal of Educational Measurement, 22, 77-105.

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

Stout, W. (1990). A new item response theory modeling approach with application to unidimensionality assessment and ability estimation. Psychometrika, 55, 293-325.

Stout, W., Nandakumar, R., Junker, B., Chang, H-H., & Steidinger, D. (1991). DIMTEST and TESTSIM [computer application software]. Urbana-Champaign: University of Illinois, Dept. of Statistics.

Wang, M. (1986, April). Fitting a unidimensional model to multidimensional item response data. Paper presented at the ONR contractors conference, Gatlinburg, TN.

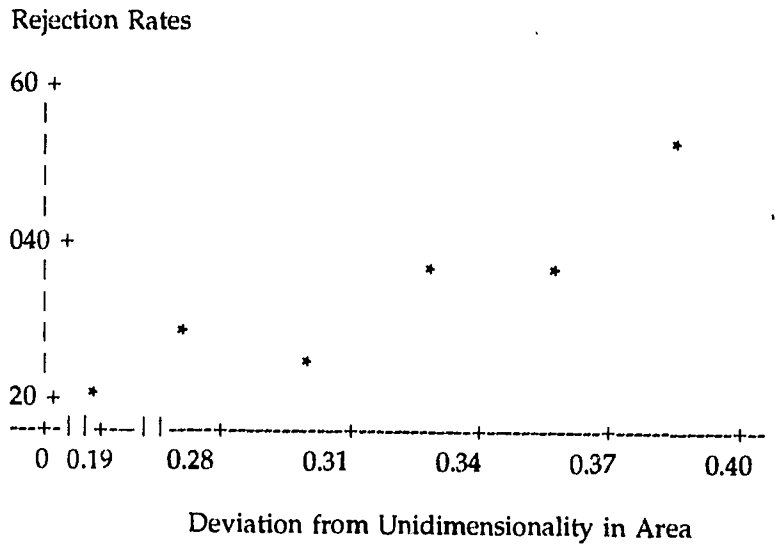

Figure 4.1

Power Curves for Deviations from Unidimensionality: Sample Size = 700, Proportion (on 2nd Dimension) = 100% and Test Length = 20

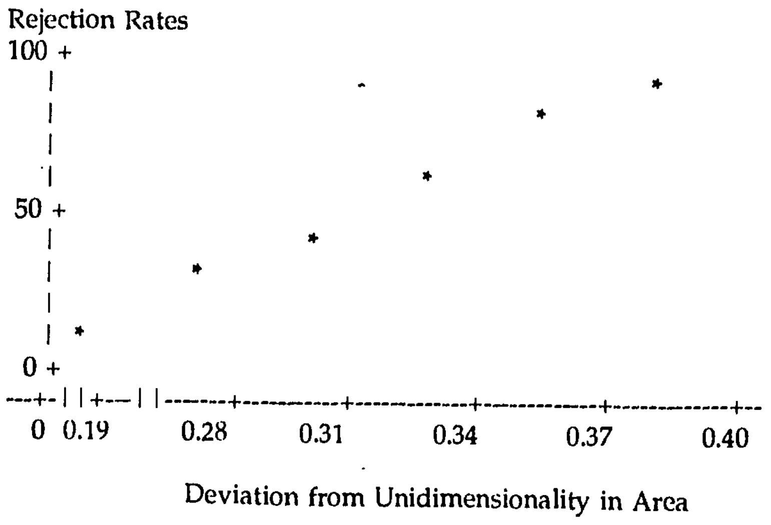

Figure 4.2

Power Curves for Deviations from Unidimensionality: Sample Size = 700, Proportion = 100% and Test Length = 40

```
Rejection Rates
    |
80 +
    |                                                                                             *
    |
60 +                                                                  *              *
    |
    |
40 +                                    *              *
    |   *
    |
20 +
---+-| |+---| |------+---------------+---------------+---------------+---------------+--
   0  0.19       0.28            0.31            0.34            0.37            0.40

                  Deviation from Unidimensionality in Area
```

Figure 4.3

Power Curves for Deviations from Unidimensionality: Sample Size = 1,500, Proportion = 100% and Test Length = 20

```
Rejection Rates
100 +                                                             *              *              *
     |
     |                                          *
     |                             *
     |
 50 +
     |   *
     |
     |
     |
  0 +
-+-| |+---| |--------+---------------+---------------+---------------+---------------+--
  0  0.19       0.28            0.31            0.34            0.37            0.40

                  Deviation from Unidimensionality in Area
```

Figure 4.4

Power Curves for Deviations from Unidimensionality: Sample Size = 1,500, Proportion = 100% and Test Length = 40

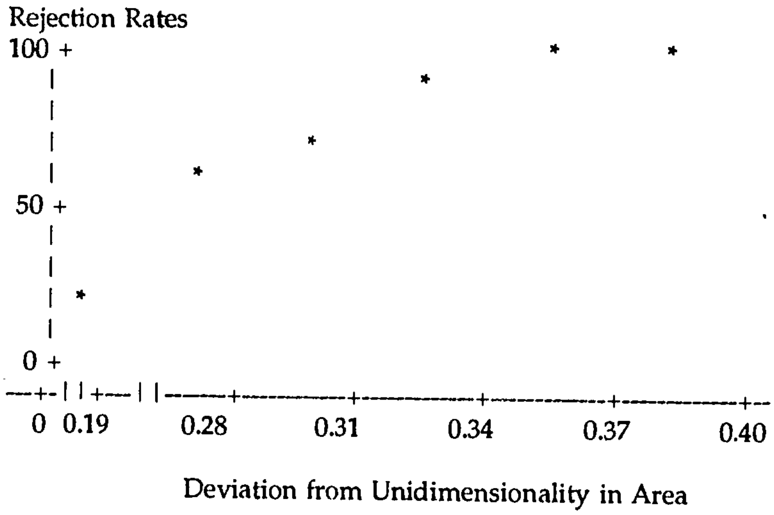

Figure 4.5

Power Curves for Deviations from Unidimensionality: Sample Size = 700, Proportion = 20% and Test Length = 20

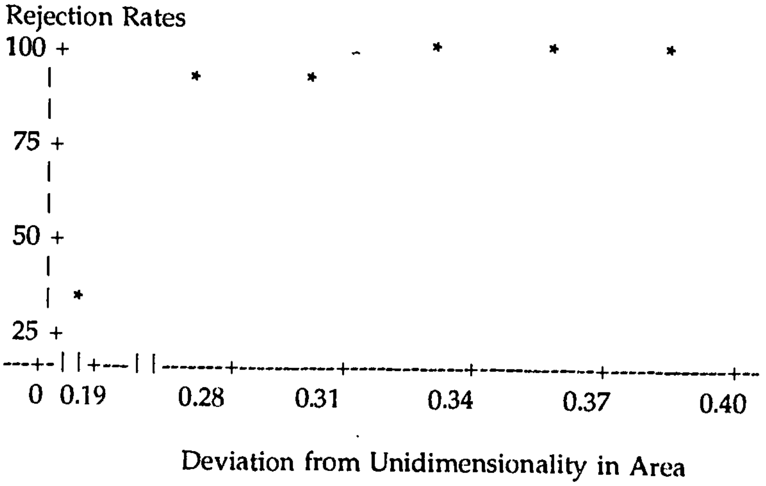

Figure 4.6

Power Curves for Deviations from Unidimensionality: Sample Size = 700, Proportion = 20% and Test Length = 40

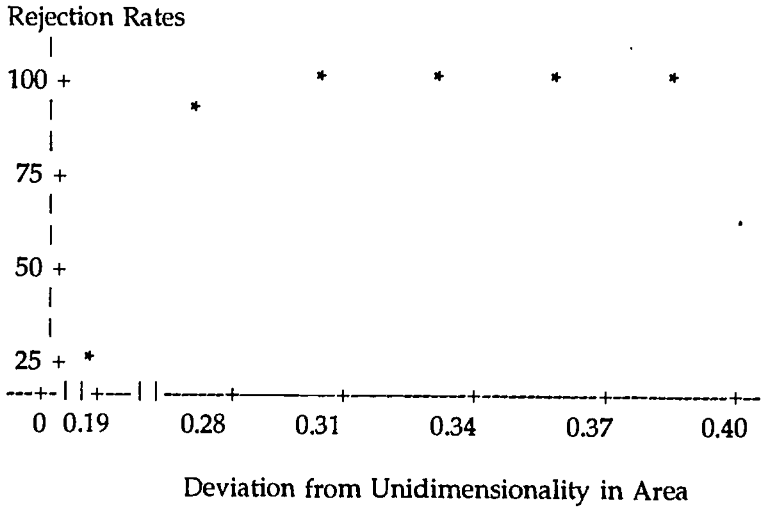

Figure 4.7

Power Curves for Deviations from Unidimensionality: Sample Size = 1,500, Proportion = 20% and Test Length = 20

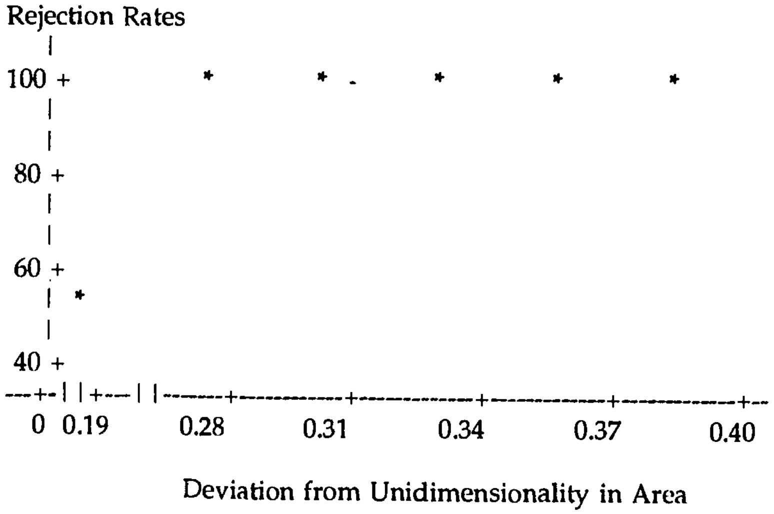

Figure 4.8

Power Curves for Deviations from Unidimensionality: Sample Size = 1,500, Proportion = 20% and Test Length = 40

```
Rejection Rates
   |
30 +
   |                                          *          *
   |
20 +           *                    *
   |                    *
   |
10 +
   |
   |  *
 0 +
--+-||+--||---+----------+----------+----------+----------+--
  0 0.19   0.28       0.31       0.34       0.37       0.40
```

Deviation from Unidimensionality in Area

Figure 4.9

Power Curves for Deviations from Unidimensionality: Sample Size = 700, Proportion = 10% and Test Length = 20

```
Rejection Rates
100 +                                           *         *
    |                                *
    |                     *
    |
    |          *
 50 +
    |
    |
    |
    | *
  0 +
---+-||+--||-------+----------+----------+----------+----------+--
   0 0.19    0.28       0.31       0.34       0.37       0.40
```

Deviation from Unidimensionality in Area

Figure 4.10

Power Curves for Deviations from Unidimensionality: Sample Size=700, Proportion = 10% and Test Length = 40

```
Rejection Rates
   |
80 +
   |
   |
60 +                                                *           *
   |
   |                                    *
40 +
   |
   |  *        *            *
20 +
---+-| |+---| |----+------------+------------+------------+------------+--
   0  0.19      0.28         0.31         0.34         0.37         0.40
```

Deviation from Unidimensionality in Area

Figure 4.11

Power Curves for Deviations from Unidimensionality: Sample Size=1,500, Proportion = 10% and Test Length = 20

```
Rejection Rates
    |
100 +                                          *           *           *
    |           *            *
    |
 75 +
    |
    |
 50 +
    |
    |  *
 25 +
---+-| |+---| |------+------------+------------+------------+------------+--
   0  0.19       0.28         0.31         0.34         0.37         0.40
```

Deviation from Unidimensionality in Area

Figure 4.12

Power Curves for Deviations from Unidimensionality: Sample Size=1,500, Proportion = 10% and Test Length = 40