Alternative Hypothesis Testing Procedures for DIMTEST

Applied Psychological Measurement
2015, Vol. 39(6) 480–493
© The Author(s) 2015
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621615577618
apm.sagepub.com

Vincent Kieftenbeld¹ and Ratna Nandakumar²

Abstract

Many commonly used item response models make the unidimensionality assumption of a single latent trait underlying the response data. The validity of this assumption needs to be tested before these models can be applied. One option is to use Stout's non-parametric hypothesis test of essential unidimensionality, which is operationalized in the DIMTEST procedure. Although generally successful, Type I error rates of this procedure are usually deflated in small samples and inflated in large samples, while power can be low in small samples. A possible cause for the unfavorable Type I error rates and power may be that estimates of the sampling distribution, bias, and standard error of the test statistic are not sufficiently accurate in finite samples. In this study, five alternative hypothesis testing procedures were formulated that replace the (asymptotically correct) approximations in the current DIMTEST procedure with computational alternatives. The performance of these procedures was investigated in two simulation studies. One of these alternative procedures, which uses a conditional covariance statistic directly in a bootstrap hypothesis test, exhibited better controlled Type I errors and higher power than the current DIMTEST procedure in most conditions. Averaged over all sample sizes and correlations between two underlying dimensions, power increased by 5 percentage points for simple structure and by 7 percentage points for approximate simple structure.

Keywords

unidimensionality, bootstrap, DIMTEST, item response theory

Many item response models currently in use for educational and psychological measurement assume that a single latent trait underlies the response data. When this assumption is violated, item and person parameter estimates obtained from unidimensional models may be confounded by multidimensionality (Ansley & Forsyth, 1985; Reckase, Carlson, Ackerman, & Spray, 1986). Confirming the (sufficient) unidimensionality of response data is therefore an important preliminary step before a unidimensional item response model can be used operationally.

Stout (1987) proposed a non-parametric hypothesis test of essential unidimensionality (a single dominant ability underlying response data) based on his operationalization of the

¹McGraw-Hill Education CTB, Monterey, CA, USA

²University of Delaware, Newark, DE, USA

Corresponding Author:

Vincent Kieftenbeld, McGraw-Hill Education CTB, 20 Ryan Ranch Road, Monterey, CA 93940-5703, Newark, DE, USA.
Email: vincent.kieftenbeld@ctb.com

unidimensionality concept in terms of conditional covariances. Codified in the DIMTEST procedure (Nandakumar & Stout, 1993; Stout, Froelich, & Gao, 2001; Stout, Nandakumar, Junker, Chang, & Steidinger, 1992), the test items are first divided into an assessment subtest (AT) and a partitioning subtest (PT) in such a way that when multidimensionality is suspected, the AT items are believed to be dimensionally distinct from the PT items. The AT items should ideally measure a single dimension. Although the selection of the AT and PT items could be done through expert opinion, this tends to be unreliable (Stout et al., 1996). In the current DIMTEST procedure, the AT and PT items are selected using an automated procedure (ATFIND) based on hierarchical cluster analysis (Froelich & Habing, 2008; Roussos, Stout, & Marden, 1998; Stout et al., 1996).

After the test items have been divided, a conditional covariance test statistic is computed using the AT items (conditioned on the total summed PT score), standardized using asymptotic variance estimates, and corrected for bias using a non-parametric bootstrap procedure (Stout et al., 2001). The final test statistic approximately has an asymptotic standard normal distribution under the null hypothesis of essential unidimensionality (Stout, 1987; Stout et al., 2001).

DIMTEST has been compared favorably with other methods, including non-linear factor analysis (Nandakumar, 1994), eigenvalues of the item covariance matrix, Yen's Q_3 statistic (DeMars, 2003), NOHARM-based procedures when guessing is present (Finch & Habing, 2007), the likelihood ratio test, and modified parallel analysis (Finch & Monahan, 2008). Although generally successful in detecting violations of unidimensionality, DIMTEST does have some limitations. Type I error rates are known to be conservative under some conditions (indicating a potential loss of power) but inflated under other conditions. For example, with a nominal level $\alpha = .05$, Socha and DeMars (2013a) reported Type I error rates below 4% in samples with 2,000 or fewer examinees, and up to 12% in samples of 4,000 examinees. Finch and Habing (2007) reported similar Type I error rates in samples of 2,000 examinees responding to a 15-item test with no guessing. Although Type I error rates usually inflate in larger samples, Finch and Monahan (2008) have observed Type I error rates of 25% in samples of 250 examinees. In general, Type I error rates tend not to be as extreme for longer tests but can still exceed the nominal level, especially in larger samples.

In smaller samples, the power of DIMTEST may fall below acceptable levels (Gessaroli & De Champlain, 1996; Pyo, 2000). Therefore, at least 750 examinees and 25 items are commonly recommended as a minimum (Stout, 1987), although DIMTEST may be applicable to fewer examinees and/or fewer items in certain conditions (Fay, 2013). More generally, it would be worthwhile to improve the power of DIMTEST to detect multidimensionality.

Purpose of the Present Study

Given the varying Type I error rates and power, it could be hypothesized that the sampling distribution of the DIMTEST statistic is not accurately approximated by its asymptotic reference distribution, $N(0, 1)$. Indeed, Socha and DeMars (2013b) concluded that when no guessing is present in the response data, the sampling distribution of Stout's test statistic is sometimes shifted to the right compared to the standard normal distribution. Instead of relying on the asymptotic normal distribution, one can empirically investigate the true distribution of the DIMTEST statistic through extensive use of bootstrapping. Currently, a non-parametric bootstrap is used only for bias correction (Stout et al., 2001), but it could also be used to estimate the finite-sample variance of the test statistic (instead of using the approximate asymptotic variance estimate) or to compute a p value from an empirical distribution (instead of using an approximate asymptotic distribution). Incorporating additional bootstrap procedures to harness

the power of sample data might provide better controlled Type I error rates while retaining or even increasing power.

The present study investigated the performance of several alternative bootstrap hypothesis testing procedures for DIMTEST. These alternative procedures were evaluated based on their Type I error rates and power. The Type I error rate and power of the current DIMTEST procedure served as the baseline for comparing alternative procedures. The main research goal was to determine the (empirical) Type I error rates and power of the alternative hypothesis testing procedures and to investigate whether these methods improve upon the current DIMTEST procedure.

The Current DIMTEST Procedure

This section explains the current DIMTEST procedure in some detail (in part to fix notation). For a detailed description, see Stout et al. (2001). In the present study, ATFIND (Roussos et al., 1998; Stout et al., 1996) was used to divide items into an assessment subset AT and a partitioning subset PT. ATFIND is based on hierarchical cluster analysis, which has been shown to be preferable over factor analysis of tetrachoric correlations (Froelich & Habing, 2008).

Denote the vector of examinee responses to item i by Y_i and the vector of total (sum) scores on the PT by R_{PT} . The k th ability subgroup consists of all examinees with $R_{PT} = k$. For each ability subgroup k , consider the sum of the covariances of the AT items within the homogeneous group, denoted by $T_{U,k}$:

$$T_{U,k} = \sum_{i \neq j \in AT} cov(Y_i, Y_j | R_{PT} = k). \quad (1)$$

All ability subgroups are combined into a single conditional covariance statistic T_U , defined as

$$T_U = \sum_{k=0}^K T_{U,k}. \quad (2)$$

When a test is essentially unidimensional, the AT and PT items measure the same dimension, and the expected value of T_U will be close to 0. However, when multidimensionality is present, AT items will be dimensionally different from PT items, and the expected value of T_U will be large.

Stout (1987) derived a formula for the asymptotic variance of $T_{U,k}$, denoted by S_k^2 . Because the computational details are not central to the present study, they are not presented here. For details, see Stout et al. (2001), in particular, Equation 19.23. An asymptotic estimate of the variance of T_U is given by

$$S^2 = \sum_{k=0}^K S_k^2. \quad (3)$$

This variance estimate is used to standardize T_U , resulting in the standardized conditional covariance statistic T_S , defined as

$$T_S = \frac{T_U}{\sqrt{S^2}}. \quad (4)$$

Although T_S does follow an asymptotic standard normal distribution (Stout, 1987), the test statistic is biased when used with finite samples, and this bias needs to be corrected before reliable p values can be computed using the standard normal distribution. Originally, Stout (1987) used a second set of AT items for bias correction, but Stout et al. (2001) demonstrated that estimating the bias using a bootstrap procedure yields higher power. This procedure estimates item response functions based on a non-parametric item response model fitted through the kernel smoothing method (Ramsay, 1991) and used to generate replicated data sets $Y^{(b)}$ for $b = 1, \dots, B$ (by default, $B = 100$). For each of the B replicated data sets, the statistic $T_S^{(b)}$ is computed ($b = 1, \dots, B$). Their mean $\bar{T}_S$ provides an estimate of the finite-sample bias of T_S :

$$\bar{T}_S = \frac{1}{B} \sum_{b=1}^B T_S^{(b)}. \quad (5)$$

Finally, Stout's DIMTEST statistic T equals

$$T = \frac{1}{\sqrt{1 + \frac{1}{B}}} (T_S - \bar{T}_S). \quad (6)$$

The denominator $\sqrt{1 + 1/B}$ is a further correction for the fact that only finitely many bootstrap samples are used (note that the denominator tends to 1 as B tends to infinity). The test statistic T has approximately an asymptotic standard normal distribution under the null hypothesis (Stout et al., 2001), and hence can be used to test the item response model assumption of essential unidimensionality.

Alternative Hypothesis Testing Procedures

The current DIMTEST procedure involves three approximations: (a) the sampling distribution of T (see Equation 6) for finite samples is approximated by $N(0, 1)$, (b) the standard error of the conditional covariance statistic T_U (Equation 2) is approximated using the asymptotic variance estimate S^2 (Equation 3), and (c) the non-zero bias of T_S in finite tests is approximated by $\bar{T}_S$ defined by Equation 5. Although these approximations are asymptotically correct, they may not be adequate in finite samples, causing Type I error rates to deviate from nominal levels. Instead of using asymptotic approximations, the sampling distribution of T (as well as T_U and T_S) and the standard error of T_U also can be estimated using bootstrap samples. Of course, this process replaces asymptotic approximations with bootstrap approximations. However, in contrast to asymptotic approximations, practitioners have the flexibility to increase the accuracy of bootstrap approximations when necessary by generating more bootstrap replications. In this study, three alternative bootstrap hypothesis testing DIMTEST procedures were developed to overcome the limitations of the three approximations listed above. In addition, two more DIMTEST procedures were developed based on robust statistics, to mitigate the possible impact of outliers among the bootstrap samples. These five alternative methods are described in the rest of this section.

Each method is based on a different test statistic, with bootstrap p values computed as follows. Given any observed test statistic t and bootstrap replications $t^{(b)}$ under the null hypothesis, a (single-tailed) p value can be computed using an empirical distribution:

$$\hat{p} = \frac{1}{B} \sum_{b=1}^B I(t^{(b)} > t), \quad (7)$$

where $I(\cdot)$ denotes the indicator function (e.g., $I(t^{(b)} > t) = 1$ if $t^{(b)} > t$ and 0, otherwise). For example, when $t = T$ from Equation 6, $T^{(b)}$ denotes the value of the T statistic computed for the b th bootstrap replication (similar notation is used throughout this article for other statistics), and a bootstrap p value for the observed DIMTEST statistic $T^{(obs)}$ can be computed from the empirical distribution $T^{(b)}$ using Equation 7. The null hypothesis is rejected when $\hat{p} < \alpha$.

Method 1

The first alternative method (M_1) replaces the standard normal distribution as a reference distribution for T by an empirical distribution. For computational reasons, the statistic T_S (Equation 4) is used instead of T (Equation 6), because computing each bootstrap replicate $T^{(b)}$ would require an additional bootstrap to estimate $\bar{T}_S^{(b)}$ for $b = 1, \dots, B$. The resulting increase in computational requirements will be prohibitive in practice. Because each $\bar{T}_S^{(b)}$ estimates the same quantity (namely, the finite sample bias of T_S), these estimates differ only due to sampling error. If it is assumed that each $\bar{T}_S^{(b)}$ is approximately equal to $\bar{T}_S$, then comparing T against its empirical distribution is equivalent to comparing T_S against its empirical distribution. This follows from the fact that bootstrap p values are invariant under order-preserving transformations. Therefore, the first alternative method M_1 compared T_S against its empirical bootstrap distribution. Note that a bootstrap p value automatically incorporates a bootstrap bias correction: Subtracting the mean of the bootstrap replicates from the observed value $T_S^{(obs)}$ as well as each bootstrap replicate $T_S^{(b)}$ would not change the bootstrap p value.

Method 2

The second alternative method (M_2) replaces the asymptotic estimate S^2 (Equation 3) with $\hat{S}^2 = \text{var}(T_U^{(b)})$, which is a bootstrap estimate of the variance of T_U . This yields a new test statistic T_S^* given by

$$T_S^* = \frac{T_U}{\sqrt{\hat{S}^2}}. \quad (8)$$

Given that T_S is biased in finite samples, it is likely that T_S^* is biased as well, and computing p values for T_S^* using $N(0, 1)$ may not be effective. The (potential) finite-sample bias can be corrected in a similar fashion as current DIMTEST procedure. Let $\bar{T}_S^*$ denote the mean of the bootstrap replications of the respective statistics, and define

$$T^* = \frac{1}{\sqrt{1 + \frac{1}{B}}} (T_S^* - \bar{T}_S^*). \quad (9)$$

The test statistic T^* , compared against the standard normal distribution, can be considered an analog of T (Equation 6), which avoids the asymptotic variance estimate. The same bootstrap samples were used to calculate the variance estimate $\hat{S}^2$ and the bias correction $\bar{T}_S^*$.

Method 3

The third alternative method (M_3) avoids the dependency on asymptotic approximations altogether. One possibility would be to compare T^* against its empirical distribution. As was the case with bootstrapping T_S instead of T , additional bootstrap replications could be avoided by

bootstrapping T_S^* instead of T^* . However, the most direct way to avoid any reliance on asymptotics is to bootstrap T_U directly. From a practical point of view, Equation 9 can be rewritten as

$$\begin{aligned} T^* &= \frac{1}{\sqrt{1 + \frac{1}{B}}} (T_S^* - \bar{T}_S^*) \\ &= \frac{1}{\sqrt{1 + \frac{1}{B}}} \left(\frac{T_U}{\sqrt{\widehat{S^2}}} - \frac{\bar{T}_U}{\sqrt{\widehat{S^2}}} \right) = \frac{1}{\sqrt{1 + \frac{1}{B}}} \left(\frac{T_U - \bar{T}_U}{\sqrt{\widehat{S^2}}} \right). \end{aligned} \quad (10)$$

Therefore, the bootstrap p value for T^* will be very close to that of T_U , again because T^* and T_U are related through an order-preserving function. From a theoretical point of view, T_U is the fundamental quantity underlying Stout's test of essential unidimensionality. As the sum of conditional covariances, T_U is a direct index of lack of unidimensionality in the response data. Hence, one could argue that the main reason for using T instead of T_U in the current DIMTEST procedure is that the former has a known (asymptotic) sampling distribution, whereas the latter does not. This precludes the use of T_U in a traditional hypothesis test. But a bootstrap hypothesis allows one to use T_U directly, without any additional asymptotic approximations. Hence, Equation 2 is the test statistic for M_3 .

Besides approximation errors introduced by using asymptotic formulas in finite samples, an alternative explanation for inflated Type I error rates in the current DIMTEST procedure could be that outliers among the bootstrap replications have a negative impact on the test statistics. This impact might be relatively large when few bootstrap samples are used ($B = 100$ seems common). To investigate this hypothesis, two additional methods were studied.

Method 4

The fourth alternative method (M_4) replaces the mean $\bar{T}_S$ (Equation 5) with the median of the bootstrap replicates $T_S^{(b)}$ as a bias correction in Equation 6. This results in a robust analog T_R of T , which can be compared against $N(0, 1)$:

$$T_R = \frac{1}{\sqrt{1 + \frac{1}{B}}} \left(T_S - \text{median}_{b=1, \dots, B} T_S^{(b)} \right). \quad (11)$$

Method 5

The fifth alternative method (M_5) replaces the variance estimate with a robust alternative. In robust statistics, 0.74 times the interquartile range (IQR) can be used as an alternative for the standard deviation (see, for example, Huynh & Meyer, 2010, for a brief explanation and an application to educational measurement). Replacing the estimate $\hat{S}$ in Equation 8 by $0.74 \times \text{IQR}(T_U)$ results in a test statistic $T_{S,R}^*$:

$$T_{S,R}^* = \frac{T_U}{0.74 \times \text{IQR}(T_U)}. \quad (12)$$

This test statistic was corrected for bias and compared against the reference distribution $N(0, 1)$ in similar fashion as T_R , resulting in a second robust analog T_R^* :

$$T_R^* = \frac{1}{\sqrt{1 + \frac{1}{B}}} \left(T_{S,R}^* - \text{median}_{b=1,\dots,B} T_{S,R}^{*(b)} \right). \tag{13}$$

The current DIMTEST procedure and the five alternative methods are summarized in Table 1.

Table 1. Current DIMTEST Procedure and Five Alternatives.

Method	Test statistic	Equation	Reference distribution	Description
M ₀	T	(6)	$N(0, 1)$	Current DIMTEST procedure
M ₁	T_S	(4)	Empirical	Replaces asymptotic standard normal distribution with empirical distribution
M ₂	T^*	(9)	$N(0, 1)$	Replaces asymptotic variance estimate with empirical variance estimate
M ₃	T_U	(2)	Empirical	No reliance on any asymptotic estimate
M ₄	T_R^*	(11)	$N(0, 1)$	Replaces mean with median in bias correction
M ₅	T_R	(13)	$N(0, 1)$	Replaces mean with median in bias correction and standard deviation with interquartile range in standard error estimate

Design of Simulation Studies

In the preceding section, five methods (M₁ through M₅) were developed as alternatives to the current DIMTEST procedure (M₀). The performance of these methods was compared in two simulation studies: one investigating empirical Type I error rates and the other power. In both simulation studies, sample size was varied ($N = 500, 750, 1,000, 1,500$, and $2,000$), while test length was fixed ($L = 30$). Shorter tests were not considered because DIMTEST’s default AT selection procedure (ATFIND) tends to not work well with shorter tests.

When using the DIMTEST procedure, the response data need to be split into two subsamples: The first subsample is used to select items for the AT and PT subtests, whereas the second subsample is used to compute the test statistic. Type I error rates inflate dramatically when this procedure is not followed (Deng, Han, & Hambleton, 2013). Most existing studies used approximately a 1:2 split, but Socha and DeMars (2013a) have recently recommended an even 1:1 split to maximize power while controlling for the Type I error rate (except for a short test with a large sample). A (pseudo) guessing parameter can also be specified in the DIMTEST program. Socha and DeMars (2013b) have suggested setting this equal to 0. The present study followed both recommendations.

For the Type I error study, response data were simulated according to the three-parameter logistic item response model (3PLM) with scaling constant $D = 1.702$ (as this scale was used in previous studies):

$$P(Y = 1|\theta, a, b, c) = c + \frac{1 - c}{1 + \exp(-Da(\theta - b))}. \tag{14}$$

Person parameters were simulated from a standard normal distribution. Item parameters were simulated as $a \sim N(1.07, 0.40)$, $b \sim N(0.58, 0.88)$, and $c \sim N(0.16, 0.05)$. Discrimination and guessing parameters were constrained to $(0, \infty)$ and $[0, 1]$, respectively. These distributions,

derived from a calibration of a verbal scholastic aptitude test (Lord, 1968), resemble item parameters used in other DIMTEST studies (Froelich & Habing, 2008; Nandakumar & Stout, 1993; Stout, 1987).

For the power study, response data were generated from a two-dimensional (compensatory) three-parameter logistic model (Reckase, 1997) given by

$$P(Y = 1 | \theta, \mathbf{a}, b, c) = c + \frac{1 - c}{1 + \exp(-D(\mathbf{a} \cdot \theta - \|\mathbf{a}\|b))}, \quad (15)$$

where $\|\mathbf{a}_i\| = \sqrt{a_{i1}^2 + a_{i2}^2}$ is the Euclidean norm of the discrimination vector. Person parameters $\theta_p = (\theta_{p1}, \theta_{p2})$ were generated from a bivariate normal distribution with means 0, standard deviations 1, and correlation $\rho = .3, .5$, or $.7$. To simulate item parameters for two-dimensional data, item parameters were first simulated as in the case of the Type I error study and then each one-dimensional discrimination parameter a was transformed into a two-dimensional vector $a_1 = a \cos \varphi$ and $a_2 = a \sin \varphi$ based on an angle φ following either (a) simple structure or (b) approximate simple structure. For a simple structure, angles for the first half of the items were set to 0° (hence, these items measured only Dimension 1), whereas angles for the second half were set to 90° (corresponding to Dimension 2). For an approximate simple structure, item angles were drawn from a uniform distribution on $[0^\circ, 30^\circ]$ for the first half of the items and from a uniform distribution on $[60^\circ, 90^\circ]$ for the second half. This methodology to simulate different dimensionality structures followed a study by Froelich and Habing (2008).

There were 5 (sample size) conditions for the Type I error study, and $5 \text{ (sample size)} \times 3 \text{ (correlation)} \times 2 \text{ (structure)} = 30$ conditions for the power study. For each condition, 500 data sets were simulated and the null hypothesis of essential unidimensionality was tested at a nominal α level of .05. For each simulated data set, $B = 99$ bootstrap replications were generated, comparable with the 100 replications used by DIMTEST as default. More bootstrap replications are usually recommended (e.g., see Efron & Tibshirani, 1993), but a pilot study by the authors indicated that more bootstrap replications did not produce different results. Despite the increasing availability of computational resources (cf. Stout et al., 2001, used 30 bootstrap samples over a decade ago), a compromise had to be found for the simulation studies. When investigating a single data set, increasing the number of bootstrap replications may be prudent.

Statistical theory implies that a bootstrap hypothesis test is exact when $\alpha(B + 1)$ is an integer and the test statistic is a pivot (Beran, 1988; Dufour & Khalaf, 2001). Although the test statistics used may not be pivots (but note that Stout's DIMTEST statistic has an asymptotical standard normal distribution and therefore is at least asymptotically pivot), the resulting empirical p values are more convenient when $B = 99$, which therefore was used instead of $B = 100$. More precise control over the number of bootstrap replications is certainly possible (see, for example, Davidson & MacKinnon, 2000), but this was not considered in the present study.

Results

Simulation Study 1: Type I Error Rates

Table 2 reports empirical Type I error rates by sample size and method. Given the nominal level of $\alpha = .05$ and $n = 500$ replications, empirical Type I error rates between 0.031 and 0.069 were deemed acceptable, based on the normal approximation to a 95% confidence interval of a binomial random variable (i.e., $\alpha \pm 1.96\sqrt{\alpha(1 - \alpha)/n}$). From this perspective, the Type I error rate of the current DIMTEST procedure was deflated for $N = 500, 750$, and $1,000$, within the expected range for $N = 1,500$, and inflated for $N = 2,000$. Moreover, there was a clear trend for

Table 2. Type I Error Rate by Sample Size and Method.

Sample size (<i>N</i>)	Method (test statistic)					
	$M_0(T)$	$M_1(T_S)$	$M_2(T^*)$	$M_3(T_U)$	$M_4(T_R)$	$M_5(T_R^*)$
500	0.016	0.044	0.092	0.070	0.022	0.110
750	0.026	0.084	0.076	0.060	0.030	0.086
1,000	0.022	0.072	0.064	0.046	0.034	0.074
1,500	0.066	0.126	0.086	0.064	0.072	0.088
2,000	0.098	0.120	0.114	0.100	0.082	0.114

Note. $\alpha = .05$. Type I error rates within interval (0.031, 0.069) in bold.

the Type I error rates of the current DIMTEST procedure to increase with increase in the sample size.

The first alternative method M_1 (bootstrap hypothesis test for T_S) had inflated Type I error rates for all sample sizes except $N = 500$. The second alternative method M_2 (bootstrap variance estimate) had inflated Type I error rates for all sample sizes except $N = 1,000$. Type I error rates for these two methods were well above the error rates of the current DIMTEST procedure and also started to inflate in comparatively smaller samples. The third alternative method M_3 (bootstrap hypothesis test for T_U) had inflated Type I error rates for $N = 500$ and $N = 2,000$, with the Type I error rates within the expected range for other sample sizes. Finally, for the two methods based on robust statistics, the fourth alternative method M_4 (robust bias correction) had a pattern of Type I error rates very similar to the current DIMTEST procedure, albeit somewhat less deflated for $N = 500, 750$, and $1,000$ and somewhat less inflated for $N = 2,000$. In other words, using the median instead of the mean in the bias correction did improve Type I error rates to some extent. The Type I error rates for the fifth alternative method M_5 (robust bias correction and standard error estimate) were inflated for all sample sizes. This mirrored the observed inflated Type I error rates for the second alternative method; using a bootstrap estimate for the standard error of T_U , whether using the standard deviation or the IQR, had a detrimental effect on error rates.

Overall, the Type I error rate seemed best controlled for M_3 , the bootstrap hypothesis test based on T_U , with observed rates (almost) within expected range when the sample size was less than 2,000 examinees. In samples of 2,000 examinees, its error rate was inflated to the same extent as the current DIMTEST procedure (approximately 10% at a nominal α level of 5%). The robust test statistic T_R (M_4) had a Type I error rate somewhat better controlled than the current DIMTEST procedure. The other alternative methods did not offer an improvement.

Simulation Study 2: Power

Tables 3 and 4 report (empirical) power to detect two-dimensional simple structure and approximate simple structure, respectively, by correlation between the two latent dimensions, sample size, and method. As can be expected, power of all methods increased when the sample size increased and decreased when the correlation between the two latent dimensions increased. Moreover, all methods had substantially more power to detect (two-dimensional) simple structure than approximate simple structure.

All methods had very high power to detect a two-dimensional simple structure when the correlation between the two underlying dimensions was low ($\rho = .3$): Power was around .95 for $N = 500$ and almost 1 for larger samples. Conversely, all methods had low power to detect a two-

Table 3. Power to Detect Two-Dimensional Simple Structure by Correlation, Sample Size, and Method.

Correlation (ρ)	Sample size (N)	Method (test statistic)					
		$M_0(T)$	$M_1(T_S)$	$M_2(T^*)$	$M_3(T_U)$	$M_4(T_R)$	$M_5(T_R^*)$
.3	500	.962	.958	.974	.968	.934	.982
.3	750	.990	.990	.998	.998	.982	.998
.3	1,000	.998	.996	.998	.998	.992	.998
.3	1,500	.994	.998	1.000	.998	.990	.998
.3	2,000	.996	.992	1.000	1.000	.986	1.000
.5	500	.764	.810	.838	.826	.696	.848
.5	750	.916	.928	.962	.962	.858	.962
.5	1,000	.948	.950	.986	.982	.900	.990
.5	1,500	.980	.974	.998	.998	.962	.998
.5	2,000	.986	.980	1.000	1.000	.958	1.000
.7	500	.234	.308	.410	.380	.224	.408
.7	750	.460	.580	.636	.606	.430	.648
.7	1,000	.664	.700	.720	.704	.576	.748
.7	1,500	.840	.843	.908	.894	.764	.912
.7	2,000	.896	.882	.980	.968	.824	.976

Note. $\alpha = .05$.

Table 4. Power to Detect Two-Dimensional Approximate Simple Structure by Correlation, Sample Size, and Method.

Correlation (ρ)	Sample size (N)	Method (test statistic)					
		$M_0(T)$	$M_1(T_S)$	$M_2(T^*)$	$M_3(T_U)$	$M_4(T_R)$	$M_5(T_R^*)$
.3	500	.364	.466	.516	.484	.334	.520
.3	750	.616	.704	.728	.692	.568	.744
.3	1,000	.764	.796	.870	.846	.714	.882
.3	1,500	.886	.900	.946	.934	.852	.946
.3	2,000	.948	.940	.976	.968	.906	.978
.5	500	.132	.252	.294	.262	.160	.312
.5	750	.240	.374	.404	.380	.236	.422
.5	1,000	.394	.482	.544	.506	.360	.540
.5	1,500	.636	.698	.706	.690	.628	.722
.5	2,000	.816	.818	.900	.882	.770	.906
.7	500	.040	.128	.110	.096	.048	.130
.7	750	.086	.172	.164	.150	.094	.178
.7	1,000	.126	.208	.194	.180	.132	.216
.7	1,500	.222	.312	.292	.276	.222	.298
.7	2,000	.364	.388	.368	.330	.302	.376

Note. $\alpha = .05$.

dimensional approximate simple structure, unless the correlation between the two dimensions was low ($\rho = .3$) and the sample size was at least 750 examinees. Averaging across sample sizes and correlations, average power rates for Methods M_0 through M_5 were respectively .84, .86, .89, .89, .81, and .90 for the case of simple structure and .44, .51, .53, .51, .42, and .54 for approximate simple structure.

Table 5. Net Change in Power.

Correlation (ρ)	Sample size (N)	Simple structure		Approximate simple structure	
		$M_3(T_U)$	$M_4(T_R)$	$M_3(T_U)$	$M_4(T_R)$
.3	500	-.048	-.034	.066	-.036
.3	750	-.026	-.012	.042	-.052
.3	1,000	-.024	-.018	.058	-.062
.3	1,500	.006	-.010	.050	-.040
.3	2,000	.002	.006	.018	-.026
.5	500	.008	-.074	.076	.022
.5	750	.012	-.062	.106	-.008
.5	1,000	.010	-.060	.088	-.046
.5	1,500	.020	-.024	.056	-.014
.5	2,000	.012	-.012	.064	-.030
.7	500	.092	-.016	.002	.002
.7	750	.112	-.034	.030	.004
.7	1,000	.016	-.100	.030	-.006
.7	1,500	.056	-.082	.056	-.006
.7	2,000	.070	-.056	-.036	-.046

Note. Net gains of 5% points or more in bold.

Because there is often a tradeoff between Type I error rates and power, these two characteristics were studied together. Alternative methods M_1 , M_2 , and M_5 displayed unacceptably large inflated Type I error rates, whereas their power rates were only comparable with M_3 . These three methods are not viable alternatives for the current DIMTEST procedure. However, alternative methods M_3 and M_4 , namely, the bootstrap hypothesis test based on T_U , and the robust bias-corrected statistic T_R were studied further, as follows. The change in power for these two methods (for a given condition) was defined as the power of the alternative method minus the power of the current DIMTEST procedure (from Tables 3 and 4). The change in the error rate was similarly computed as the Type I error rate of the alternative method minus the Type I error rate of the current DIMTEST procedure (from Table 2). The difference between the change in power and the change in error rate is the net change in power for the alternative method, presented in Table 5.

The net change in power of the robust method, M_4 , is almost always negative. In other words, although using the median in the bias correction provides better control of the Type I error rate, this results in a net loss of power relative to the current DIMTEST procedure. Moreover, the net change in power for M_4 is almost always smaller than the net change in power for the T_U -based method, M_3 . Hence, M_3 is preferable over M_4 , and is the most promising alternative to the current DIMTEST procedure.

Discussion

Stout's hypothesis test for essential unidimensionality as codified in the current DIMTEST procedure uses asymptotic approximations for estimates of the sampling distribution, finite-sample bias, and standard error of the test statistic. Five alternative hypothesis testing procedures were formulated in this study that replace one or more of these approximations with computational alternatives. The Type I error rates and power of the alternative methods were compared against the current DIMTEST procedure. Four methods either failed to control Type I error rates or lost power to the current DIMTEST procedure. The remaining method, M_3 , which uses

the conditional covariance statistic T_U directly in a bootstrap hypothesis test, does offer an improvement over the current DIMTEST procedure. Henceforth, Method M_3 will be referred to as DIMTEST- T_U .

In the Type I error study, DIMTEST- T_U had Type I error rates conforming better to the expected range than the current DIMTEST, provided the sample size is less than 2,000. With 2,000 examinees, its Type I error rate was no more inflated than the current DIMTEST. In the power study, DIMTEST- T_U had higher power than the current DIMTEST procedure at almost every level of correlation and sample size. Averaged over all sample sizes and correlations between the two underlying dimensions, the gain in power was 5 percentage points for simple structure and 7 percentage points for approximate simple structure. Over all conditions, the average net change in power (i.e., the increase in power minus the increase in Type I error rate) was 3.4% points ($SD = 4.1$). The net change was 5 percentage points or more in 13 out of 30 conditions (43%). When the sample size is small ($N \leq 1,000$) and the multidimensionality is strong (simple structure with $\rho = .3$), DIMTEST- T_U does have a negative net change in power relative to the current DIMTEST procedure ($M = -3.3$, $SD = 1.3$ percentage points). This is because the Type I error rate of the current DIMTEST procedure under these conditions is below the nominal level (2.6 percent or less), while power for both methods is almost 1. Nevertheless, the Type I error rates for DIMTEST- T_U are within the expected range and its power is not practically compromised.

The new DIMTEST- T_U clearly outperforms the current DIMTEST when the multidimensionality is reasonably difficult (but not too difficult) to detect: simple structure with high correlation ($\rho = .7$) and approximate simple structure with medium correlation ($\rho = .5$). Although the gains under these conditions varied over a relatively wide range (between 1.6 and 11.2 percent points), the average net change in power for these conditions was 7.4 percentage points ($SD = 2.9$). Improvements in detecting violations of non-unidimensionality under these conditions may be practically important, as DIMTEST is often used to confirm the unidimensionality of a scale constructed to have a single underlying latent trait. In these applications, response data are likely not clearly two dimensional, and detecting moderate violations of unidimensionality is the practical focus.

In summary, DIMTEST- T_U not only controls the Type I error rate better than the current DIMTEST procedure but generally also increases power. It requires the same number of bootstrap samples as the current DIMTEST procedure. From a theoretical perspective, T_U (Equation 1) lies at the root of the definition of dimensionality and forms the foundation for the DIMTEST procedure: Essential unidimensionality is violated if the conditional covariance is not “small.” From a practical perspective, DIMTEST- T_U provides a genuine improvement without a real drawback over the current DIMTEST procedure, replacing asymptotics with modern computational statistics (e.g., bootstrap hypothesis test).

Other statistical methods for testing dimensionality also avoid dependency on non-empirical reference distributions, in particular, posterior predictive model checking (PPMC; Levy, Mislevy, & Sinharay, 2009; Sinharay, Johnson, & Stern, 2006). The methodology of PPMC and bootstrap hypothesis testing is very similar, in that a p value of a test statistic (usually called a discrepancy measure in the context of PPMC) is derived from an empirical distribution estimated from replicated data sets. Whereas the replicated data sets in PPMC are generated from a parametric model (and uncertainty in model parameter estimates is taken into account), these data are generated through a non-parametric model in the DIMTEST- T_U procedure. This can be considered an advantage (no assumptions are made regarding the underlying item response model) or a disadvantage (power may be lost when the parametric model is correct). A comparison of parametric and non-parametric empirical methods to detect the lack of

unidimensionality would be valuable, especially if it considered performance of methods under model–data misfit.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

- Ansley, R. A., & Forsyth, T. N. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement*, 9, 37-48.
- Beran, R. (1988). Prepivoting test statistics: A bootstrap view of asymptotic refinements. *Journal of the American Statistical Association*, 83, 687-697.
- Davidson, R., & MacKinnon, J. G. (2000). Bootstrap tests: How many bootstraps? *Econometric Reviews*, 19, 55-68.
- DeMars, C. E. (2003). Detecting multidimensionality due to curricular differences. *Journal of Educational Measurement*, 40, 29-51.
- Deng, N., Han, K. T., & Hambleton, R. K. (2013). A review of DIMPACK version 1.0: Conditional covariance-based test dimensionality analysis package. *Applied Psychological Measurement*, 37, 162-172.
- Dufour, J. M., & Khalaf, L. (2001). Monte Carlo test methods in econometrics. In B. Baltagi (Ed.), *A companion to theoretical econometrics* (pp. 494-519). Oxford, UK: Blackwell.
- Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York, NY: Chapman & Hall.
- Fay, D. M. (2013, April). *An evaluation of the performance of DIMTEST with conditional covariance-based subset selection methods with small samples and short tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
- Finch, H., & Habing, B. (2007). Performance of DIMTEST- and NOHARM-based statistics for testing unidimensionality. *Applied Psychological Measurement*, 31, 292-307.
- Finch, H., & Monahan, P. (2008). A bootstrap generalization of modified parallel analysis for IRT dimensionality assessment. *Applied Measurement in Education*, 21, 119-140.
- Froelich, A. G., & Habing, B. (2008). Conditional covariance-based subset selection for DIMTEST. *Applied Psychological Measurement*, 32, 138-155.
- Gessaroli, M. E., & De Champlain, A. F. (1996). Using an approximate chi-square statistic to test the number of dimensions underlying responses to a set of items. *Journal of Educational Measurement*, 33, 157-179.
- Huynh, H., & Meyer, P. (2010). Use of robust z in detecting unstable items in item response theory models. *Practical Assessment, Research & Evaluation*, 15, 1-7.
- Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement*, 33, 519-537.
- Lord, F. M. (1968). An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement*, 28, 989-1020.
- Nandakumar, R. (1994). Assessing dimensionality of a set of item responses: Comparison of different approaches. *Journal of Educational Measurement*, 31, 17-35.
- Nandakumar, R., & Stout, W. F. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational and Behavioral Statistics*, 18, 41-68.
- Pyo, K. H. (2000, April). *Assessing dimensionality of a set of language test data with simulated and real test data*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

- Ramsay, J. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, 56, 611-630.
- Reckase, M. D. (1997). A linear logistic model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271-286). New York, NY: Springer.
- Reckase, M. D., Carlson, J. E., Ackerman, T. A., & Spray, J. A. (1986, June). *The interpretation of unidimensional IRT parameters when estimated from multidimensional data*. Paper presented at the annual meeting of the Psychometric Society, Toronto, Ontario, Canada.
- Roussos, L., Stout, W., & Marden, J. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. *Journal of Educational Measurement*, 43, 215-243.
- Sinharay, S., Johnson, M., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, 30, 298-321.
- Socha, A., & DeMars, C. E. (2013a). An investigation of sample size splitting on ATFIND and DIMTEST. *Educational and Psychological Measurement*, 73, 631-647.
- Socha, A., & DeMars, C. E. (2013b). A note on specifying the guessing parameter in ATFIND and DIMTEST. *Applied Psychological Measurement*, 37, 87-92.
- Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika*, 52, 589-617.
- Stout, W. F., Froelich, A. G., & Gao, F. (2001). Using resampling methods to produce improved DIMTEST procedure. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays in item response theory* (pp. 357-375). New York, NY: Springer.
- Stout, W. F., Habing, B., Douglas, J., Kim, H. R., Roussos, L. A., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement*, 20, 331-354.
- Stout, W. F., Nandakumar, R., Junker, B., Chang, H.-H., & Steidinger, D. (1992). DIMTEST: A Fortran program for assessing dimensionality of binary item responses. *Applied Psychological Measurement*, 16, 236.