An Investigation of Sample Size Splitting on ATFIND and DIMTEST

Alan Socha¹ and Christine E. DeMars¹

Abstract

Modeling multidimensional test data with a unidimensional model can result in serious statistical errors, such as bias in item parameter estimates. Many methods exist for assessing the dimensionality of a test. The current study focused on DIMTEST. Using simulated data, the effects of sample size splitting for use with the ATFIND procedure for empirically deriving a subtest composed of items that potentially measure a second dimension versus DIMTEST for assessing whether this subtest represents a second dimension were investigated. Conditions explored included proportion of sample used for ATFIND, sample size, test length, interability correlations, test structure, and distribution of item difficulties. Overall, it appears that DIMTEST has Type I error rates near the nominal rate and good power in detecting multidimensionality, although Type I error inflation is observed for larger sample sizes. Results suggest that a 50/50 split maximizes power and keeps the Type I error rate below the nominal level unless the test is short and the sample is large. A 75/25 split controls Type I error better for short tests and large samples.

Keywords

ATFIND, DIMTEST, item response theory, unidimensionality, multidimensionality

The assumption of unidimensionality is useful if one wants to report a single score for a test, for writing items, and for developing statistical procedures for test development, scoring, and evaluation (Roussos & Ozbek, 2006). Unidimensionality is an assumption of the standard one-, two-, and three-parameter item response theory (IRT) models and some polytomous IRT models, including the graded response,
partial credit, and nominal response models. All of these models estimate a unidimensional latent variable/construct (θ) representing ability, demonstrated proficiency, or latent trait (the term ability will be used throughout the current article to represent all these). Serious statistical errors in evaluating individuals and in achieving various testing objectives can result from unjustified use of these models (Nandakumar, 1991). For instance, a violation of unidimensionality can result in bias of item parameter and examinee ability estimates. Reckase (1985) and Ansley and Forsyth (1985) found that treating multidimensional data as unidimensional can result in upwardly biased item difficulty estimates, and discrimination and ability estimates that are the average of those underlying the data. Treating multidimensional data as unidimensional can also cause the misidentification of differential item functioning (Ackerman, 1992; Camilli, 1992; Oshima & Miller, 1992). That is, misspecification of the latent ability space, such as measuring multiple abilities as a single ability, can result in what appears to be differential item functioning between different groups. Furthermore, modeling one underlying ability dimension when there are, in actuality, multiple dimensions, results in a loss of information.

Having multidimensional data does not always imply that multidimensional models should be used. Conditions exist in which multidimensional test data can be appropriately modeled as unidimensional. For example, even though most items on a test may be multidimensional, the strength of each dimension may not be strong and a unidimensional model may still be appropriate (Reckase, 1985). That is, there may be more than one dimension but only one dimension has a high discrimination for a particular item. Additionally, tests that contain items that measure the same weighted composite of multiple dimensions may also still meet the requirements of a unidimensional model, despite requiring several demonstrated abilities to answer the items correctly (Reckase, Ackerman, & Carlson, 1988). A third example is when examinees only vary in their level of one of the abilities (Ackerman, 1994). Mathematically, the data will be unidimensional in these latter two examples, even though the data are conceptually multidimensional.

Regardless, it is still important to investigate the assumption of unidimensionality. A unidimensional latent variable is not always appropriate, as sets of test items may measure several abilities instead of one (Ackerman, 1994; Reckase, 1985; Reckase et al., 1988). Various procedures, such as McDonald’s nonlinear factor analyses (McDonald, 1967), Holland and Rosenbaum’s conditional association approach, (Holland & Rosenbaum, 1986) and Nandakumar and Stout’s (1993) DIMTEST procedure, can be used to test the assumption of unidimensionality (Ackerman, 1994). Residual analysis is also useful in determining whether too few dimensions are specified in the model (Reckase, 1997). When unidimensionality is rejected, other options include testlet scoring, multidimensional compensatory logistic modeling, and separating the test into several essentially unidimensional subtests.

The focus of the current study is the DIMTEST procedure (DIMPACK Version 1, 2006). Stout (1987) gave three situations in which a statistical test for unidimensionality, such as DIMTEST, can be useful:
(a) avoid contamination of a test by another trait, (b) indicate whether a test (even if the items it contains are appropriate for its purpose) is measuring two (or more) traits and hence should be split into two (or more) subtests for analysis and interpretation, and (c) carry out as a preliminary step prior to the use of any of the standard (unidimensional) latent trait methodologies. (p. 590)

A few studies have compared the DIMTEST procedure with other procedures for detecting multidimensionality. Finch and Habing (2007) compared DIMTEST with the NOHARM-based procedures of $\chi^2_{G/D}$, ALR, and $T_s$ using data generated with and without guessing and found that DIMTEST had more accurate Type I error rates than the NOHARM-based procedures when guessing was simulated. These procedures controlled Type I error rates well when there was no correct guessing in the data, and the NOHARM-based procedures had somewhat better power, except when the interability correlation ($r = .95$) was high or the ability distribution was skewed. Finch and Monahan (2008) compared DIMTEST, the likelihood ratio test, and a modified parallel analysis and found that DIMTEST had lower Type I error rates and comparable power when there were a larger number of examinees (i.e., $N > 250$).

Nandakumar (1994) compared DIMTEST with Holland and Rosenbaum’s approach and nonlinear factor analysis and found that DIMTEST had higher power in detecting multidimensionality. Nandakumar (1994) also found that DIMTEST was the most efficient in regard to computational time. Finally, DeMars (2003) used DIMTEST, the -2LL test for the difference between a one-factor and two-factor IRT model, the eigenvalues of the item covariance matrix, and Yen’s Q3 to assess multidimensionality for complex items measuring one general trait plus a secondary trait. DIMTEST had accurate Type I error and generally good power (DeMars, 2003). In contrast the -2LL test of the difference between one-factor and two-factor had extreme alpha inflation. The eigenvalues and the mean Q3 were not very sensitive to multidimensionality.

DIMTEST is a non-IRT procedure that compares an Assessment Subtest (AT) composed of items that potentially measure a second dimension with a Partitioning Subtest (PT) composed of all other items on the test. The AT can be defined theoretically, by choosing the items suspected of forming the secondary dimension, or empirically, through the use of statistical procedures. When the AT is chosen empirically, to reduce capitalizing on chance, one subset of the sample is used to select the AT and a different subset is used to test whether the AT measures a significantly different second dimension. This study will explore how the sample should be split to minimize Type I error and maximize power. For example, should a smaller sample be used to select the AT, leaving a large sample for the statistical significance test?

ATFIND, which is also a component of DIMPACK but a separate procedure from DIMTEST, uses an agglomerative hierarchical cluster analysis (HCA/CCPROX) procedure and is one empirical method for deriving the AT. HCA/CCPROX is a non-parametric procedure where each item starts out as its own cluster in the first stage. At each successive stage the two closest clusters are joined until the final stage where all items are in a single cluster (Roussos, Stout, & Marden, 1998; Stout et al., 1996).
HCA/CCPROX uses the unweighted pair-group method of averages in determining
the proximity of the clusters. In other words, proximity of the clusters is determined
via the average covariance between items, conditioned on the remaining items. The
higher the conditional covariance, the closer the clusters are in proximity, with posi-
tive covariances representing clusters that are close together and negative covariances
representing clusters that are far apart. A Dimensionality Evaluation to Enumerate
Contributing Traits (DETECT) index, representing how dimensionally distinct a set
of clusters are, is calculated at each stage between each cluster of at least four items
and the remaining items not in that cluster. The amount of multidimensionality is
indicated by the maximum DETECT value, so the cluster with the highest DETECT
index is chosen as the AT (Roussos & Ozbek, 2006; Zhang & Stout, 1999).

The DIMTEST procedure then computes the pairwise item covariances of the AT,
conditioned on score on the PT, to see if they differ from zero (Stout et al., 1996).
The covariances should be near zero if the items do not share a secondary dimension
because examinees within each score group will have the same estimated score on
the primary dimension. The procedure starts by computing the conditional covar-
iances in the AT among examinees with the same raw score on the PT. The within-
group covariance (i.e., the variance of the AT raw scores minus the sum of the item
variances) is computed and standardized. The results are summed across score groups
and divided by the square root of the number of score groups, which yields

\[ T_L = \frac{1}{\sqrt{K}} \sum_{k=1}^{K} \frac{T_{L,k}}{S_k}, \]

where \( T_{L,k} \) is the sum of the interitem covariances within score group \( k \) (i.e., \( T_{L,k} \)

is the sum of \( \text{cov}(U_i, U_l | \theta_{PT}) \) over all pairs \( i, l \) in AT where \( U_i \) is the score on item \( i \)

and \( \theta_{PT} \) is the PT subtest score) and \( S_k \) is an estimate of the variance of \( T_{L,k} \) (Stout,
Froelich, & Gao, 2001).

\( T_L \) is positively biased, especially for short tests (Stout, 1987; Stout et al., 2001).
Early versions of DIMTEST required three subtests instead of two to correct for this:
AT1, AT2, and PT. AT1 items are those hypothesized to measure a different dimen-
sion than the remaining items on the test. AT2, containing items that are selected to
be similar in difficulties to AT1 items but measuring the same dimension as PT, was
used to correct for the bias in \( T_L \). AT2 represented the positive bias expected in AT1
under the assumption of unidimensionality. \( T_G \), a difference statistic, was calculated
for AT2 and used to correct the positive bias. Early versions of DIMTEST also had
inflated rejection rates for larger samples when both guessing and highly discrimina-
ting items (i.e., \( a > 1.0 \)) were present (Nandakumar & Stout, 1993). More appropriate
methods for selecting the AT addressed this issue, however (Nandakumar & Stout,
1993). This method of DIMTEST that used AT2 to correct for positive bias usually
had Type I error rates near the nominal rate (i.e., \( \alpha = .05 \)) and good power to detect
multidimensionality (Stout, 1987), but had the limitation that it sacrificed a portion
of the test to correct for bias.
The current version of DIMTEST addressed these limitations by using a simulated resampling method to correct for the statistical bias, thus eliminating the need for AT2 (Stout et al., 2001). This resampling method starts by estimating nonparametric item characteristic curves, formed by grouping examinees by raw score on the PT and calculating the proportion of group members who answer the item correctly. These curves are smoothed via Ramsay’s method of kernel-smoothing (Stout et al., 2001). Data are then simulated based on these item characteristic curves. These simulated data are unidimensional because they are conditional only on the PT.

This simulation process is repeated a number of times and $T_G$ is calculated for each replication, using Equation (1) with the simulated data. $T_G$ should contain the same bias as $T_L$ but is based on unidimensional data and thus can be used to correct for the positive bias. Thus, AT2’s purpose is served without the need to sacrifice a portion of the test. This resampling method has more statistical power than the original DIMTEST version and is designed to maintain the nominal $\alpha$ level and have power in detecting multidimensionality when the AT and PT are chosen appropriately (Finch & Habing, 2007). The final statistic is a standardized $T$ statistic:

$$T = \frac{T_L - T_G}{\sqrt{1 + 1/N}}.$$  \hspace{1cm} (2)

$T$ has an asymptotically standard normal distribution as the number of items and examinees increases to infinity (Finch & Habing, 2007). It has been shown that DIMTEST has an average Type I error rate near the nominal rate of $\alpha = .05$ and high power in detecting multidimensionality (Finch & Habing, 2007; Nandakumar, 1991; Nandakumar & Stout, 1993; Stout et al., 2001). Finch and Habing (2007) found that when tests are short (i.e., 15 items) and sample size is around 1,000 or 2,000, DIMTEST is more likely to have a higher Type I error rate. In these conditions and with item parameters generated based on the SAT test and under varying conditions of skewness of the ability distributions, Type I error rates ranged from .055 to .132 for data generated from the 2PL model and from .052 to .089 for 3PL data. In all other conditions, Finch and Habing (2007) found that DIMTEST was close to maintaining the nominal .05 error rate.

DIMTEST is advantageous because computational requirements are low, it is supported by theory and simulations, and it is based on a nonparametric IRT model (Stout, 1987). Several studies have assessed DIMTEST and found DIMTEST to be effective (DeMars, 2003; Finch & Habing, 2007; Finch & Monahan, 2008; Froelich & Habing, 2008; Hattie, Krakowski, Rogers, & Swaminathan, 1996; Nandakumar, 1994; Stout, 1987). Several other recent studies have used this procedure to investigate the dimensionality of scales (e.g., Hsieh, 2010; Jasper, 2010; Tourón, Lizasoain, & Joaristi, 2012). Unfortunately, there have been no studies that have investigated the effects of splitting the sample on deriving the AT versus testing the AT against the PT. Furthermore, the proportion of the sample used for ATFIND versus that used for DIMTEST has not been consistent throughout the literature and is usually not accompanied with a rationale. In deriving the AT, for example, Finch and Habing (2007)
and Zhang and Stout (1999) used one half of their sample, Froelich and Habing (2008) used one third of their sample, and Nandakumar and Stout (1993), Stout (1987), and Stout et al. (2001) used one third and one fourth of their sample.

Whereas ATFIND is often used to derive the AT used by DIMTEST, these are separate procedures, and increasing the proportion of the sample used for either can affect the overall power for detecting multidimensionality. Both procedures are included in DIMPACK, but ATFIND is optional—the user could choose the AT by other methods prior to running DIMTEST. It is the DIMTEST procedure that provides the statistical significance test of violation of unidimensionality. The current study will investigate how the percentage of the sample that is used for ATFIND versus DIMTEST affects the Type I error and power of the procedures. Increasing the proportion of the sample size for ATFIND instead of DIMTEST should lead to a better choice of AT and thus increase the power, but increasing the proportion of the sample size for DIMTEST instead of ATFIND increases the power of the statistical significance test. No study to date has investigated which of these procedures results in the greatest gain in power without sacrificing Type I error for the same sample. Therefore, no predictions are made concerning the proportion of sample used for ATFIND and its relationship to Type I error and power because it is not clear where the greatest proportion of the sample will best boost the total power.

This study will also compare the power of DIMTEST when the multidimensionality follows simple structure or complex structure. Given DETECT’s ability to determine the number of dimensions is greater when there is simple structure (Finch, Stage, & Monahan, 2008; Gierl, Leighton, & Tan, 2006; Stout et al., 1996) and the similarity of the conditional covariance procedure used in DIMTEST to that in DETECT, it is hypothesized that power will be greater when the items have simple structure rather than complex structure. It is also anticipated that power will increase as the interability correlation decreases.

**Method**

One- and two-dimensional dichotomous test data were randomly generated using PROC IML in SAS, Version 9.2 (SAS Institute, Inc., Cary, NC). The unidimensional data followed the three-parameter model (see Equation 3) and the two-dimensional data followed the two-dimensional compensatory three-parameter model (see Equation 4).

\[
P(x_i = 1) = c_i + \frac{1 - c_i}{1 + \exp\{-1.7a_i(\theta - b_i)\}}. \quad (3)
\]

\[
P(x_i = 1) = c_i + \frac{1 - c_i}{1 + \exp\{-1.7[a_{1i}(\theta_1 - b_i) + a_{2i}(\theta_2 - b_i)]\}}. \quad (4)
\]

Previous studies (Finch & Habing, 2007; Froelich & Habing, 2008; Nandakumar, 1991; Nandakumar & Stout, 1993; Seraphine, 2000; Stout, 1987; Walker, Azen, &
Schmitt, 2006) had much variation in their parameters. For example, means of item discrimination parameters ranged from 0.2 to 1.46 (SDs ranging from 0.06 to 0.70), means of difficulty parameters ranged from 0.92 to 0.58 (SDs ranging from 0.37 to 0.96), and means of guessing parameters ranged from 0.00 to 0.36 (SDs ranging from 0.00 to 0.06). Most abilities were generated from the standard normal distribution (see Froelich & Habing, 2008; Nandakumar, 1991; Nandakumar & Seraphine, 2000; Stout, 1993; Stout et al., 2001; Walker et al., 2006). The simulation parameters for the current study were chosen to resemble previous research. Some of these studies based their parameters on actual data (e.g., Finch & Habing, 2007; Froelich & Habing, 2008; Nandakumar, 1991; Nandakumar & Stout, 1993; Walker et al., 2006).

Item discrimination parameters \( (a_{i1}, a_{i2}) \) were generated from a lognormal distribution with a mean of 0 and standard deviation of 0.5, difficulty parameters \( (b_i) \) from a normal distribution with a mean of 0 and standard deviation of 0.6 (values in magnitude greater than 2 were regenerated, as these items are too easy or too hard and would likely be thrown out in any large-scale testing situation), and guessing parameters \( (c_i) \) from a uniform distribution ranging from 0.0 to 0.2. Also, data sets were generated with these same parameters, with the exception that a standard deviation of 2 was used for the difficulty parameters; because values greater than 2 in absolute value were replaced with new random draws, this yielded a platykurtic distribution of item difficulties. Such a distribution may be evident in tests where the concern is to get more information on examinees at the low and high ends of the ability distribution. Abilities \( (\theta, \theta_1, \text{and} \theta_2) \) were generated from a standard normal distribution. The examinee was considered to have answered an item correct (i.e., score of 1) if \( P(x_i = 1) \) was greater than a random number generated from a uniform distribution from 0 to 1, otherwise, the examinee was considered to have gotten the item wrong (i.e., score of 0).

The two-dimensional data were generated under the conditions of simple structure, with each item measuring only one dimension (i.e., each item had a discrimination parameter for only one dimension, with each dimension containing half the items), and complex structure, where items were free to measure either dimension or both (i.e., the items had discrimination parameters for each dimension). The discrimination parameters were drawn independently for each dimension, so some items might have high discrimination parameters for both dimensions, others might have low or moderate discriminations for both, and some might have a high value for one discrimination parameter and a low value for the other. For tests with simple structure, the HCA/CCPROX procedure will reproduce that structure as one of its clusterings (Stout et al., 1996). Conditions were picked to resemble previous research. Previous studies specified interability correlations ranging from .00 to .95, test lengths ranging from 15 to 101, and sample sizes ranging from 750 to 20,000 (see Finch & Habing, 2007; Froelich & Habing, 2008; Nandakumar, 1991; Nandakumar & Stout, 1993; Seraphine, 2000; Stout, 1987; Stout et al., 1996; Stout et al., 2001; Walker et al., 2006; Zhang & Stout, 1999). One- and two-dimensional data were crossed with dimensional structure (only for the two-dimensional data), three
interability correlation conditions (.00, .35, .70; only for the two-dimensional data), three test lengths (20, 40, and 60 items), and four sample sizes (500, 1,000, 2,000, and 4,000).

A total of 1,000 data sets were generated for each condition. Nonparametric Dimensionality Assessment Package (DIMPACK Version 1.0, 2006) containing ATFIND Version 1.3, DIMTEST Version 2.1, and DETECT Version 2.1 was used to derive the ATs and test the hypothesis of unidimensionality. Because the guessing parameters were generated from a uniform distribution with a mean of .10, .10 was specified as the guessing parameter in both ATFIND and DIMTEST. ATFIND and DIMTEST use this parameter to estimate the lowest score that can be reliably used, thus slightly reducing extraneous noise (see DIMPACK Version 1.0 documentation). An alpha level of .05 was used for rejecting the null hypothesis of unidimensionality. Different percentage splits of the sample size were specified for ATFIND versus DIMTEST for each replication. Two splits, 25/75 and 50/50, were based on prior studies, whereas one split, 75/25, was selected to test whether there is greater power in detecting multidimensionality as a result of using more data to select the AT.

Results and Discussion

Type I Error

Type I error rates, that is, the proportion of unidimensional data sets where DIMTEST found multidimensionality, for the conditions where the difficulties were generated from a normal distribution with a mean of 0 and standard deviation of 0.6 are graphed in Figure 1. Most of the relationships in this graph are very clear and intuitively make sense. First, Type I error tends to decrease as the number of items on the test increases. Type I error rates increase with sample size, though, which is not an expected property of an accurate statistical hypothesis testing procedure. It appears that Type I error inflation occurs only with extremely large sample sizes (N = 4,000) with small tests (20 items). In all other conditions the Type I error rate is too conservative, falling below the nominal .05 rate.

Type I error rate was also examined across various sample size splits between the ATFIND and DIMTEST procedures. Type I error rates between various sample splits were also similar. Here, a 75/25 split resulted in the lowest Type I error rate and a 25/75 split resulted in the highest Type I error rate, when the sample size was large. Results were similar for conditions where the difficulties were generated from a standard normal distribution with the exception that Type I error was higher across all conditions except when sample size was small (i.e., 500 and 1,000). When sample size was large, the Type I error rate was higher by as much as .12 (N = 4,000, one fourth of sample used for ATFIND and three fourths of sample used for DIMTEST).
Figure 1. Type I error rates
Power, that is, proportion of replications correctly classified as multidimensional, is graphed in Figures 2 and 3 for multidimensional tests with simple structure and multidimensional tests with complex structure, respectively. These graphs pertain to the conditions where the difficulties were generated from a normal distribution with a mean of 0 and standard deviation of 0.6. As would be expected, power tended to increase as sample size increased and as the number of items on the test increased.

Power tended to decrease as the magnitude of the interability correlation of the multidimensional tests increased. Tests with a .70 correlation had the lowest power. This result is to be expected, because the dimensions become closer to measuring the same construct as the correlation between them increases. This follows Reckase’s
statement that a unidimensional model may be appropriate if the strength of the dimensions is not strong. It is consistent with the findings of Finch and Habing (2007). Also, power was higher for tests with simple structure than for tests with complex structure, which fits with previous findings that the DETECT index is less accurate in determining the structure of the data when structure is complex and the interability correlation is high (Finch et al., 2008; Gierl et al., 2006). In fact, the conditions for tests with simple structure that had a .70 interability correlation had similar power to conditions for tests with complex structure that had a .00 interability correlation. In some conditions, such as when the interability correlation was high and the items exhibited complex structure, 60-item tests were necessary to produce acceptable power.
Power across most sample size splits between the ATFIND and DIMTEST procedures was similar. In many conditions, power was virtually 1 regardless of the sample size split. In other conditions, power was highest when a 50/50 split of the sample was used for each procedure and lowest when a 75/25 split of the sample was used for ATFIND and DIMTEST, respectively. The biggest difference in power across sample splits occurred when there was a small sample size, fewer test items, the dimensions were highly correlated, and the dimensions had a complex structure. Results were similar for conditions where the difficulties were generated from a standard normal distribution with the exception that power was lower.

**Distribution of the Test Statistic**

Because $T$ is supposed to have an asymptotically standard normal distribution (Finch & Habing, 2007), the $T$ distribution, $T_L$ distribution, and standard normal curve were graphed (see Figure 4) for the null condition (i.e., the unidimensional test conditions). These distributions should indicate whether the $T$ statistic is truly asymptotically normal and why Type I error inflation may be occurring. The $T$ distribution is shifted slightly to the right of the standard normal curve for every condition, that is, the mean is greater than zero. Looking at Figure 4, one can see that $T$ corrects for much of the bias in $T_L$ but the correction is not entirely successful at bringing $T$ to a standard normal distribution. Pertaining to the variable of interest, the proportion of the sample used for ATFIND and DIMTEST, the $T$ distribution shifts less to the right (becoming closer to the standard normal distribution) as the proportion of the sample used for DIMTEST decreases. Thus, it appears that, to reduce Type I error, less of the sample should be used in in DIMTEST. However, this obviously will reduce power.

**Limitations and Future Directions**

Several limitations and directions should be noted. First, the current study used simulated data. The ideal conditions of the current study will not always occur in practice. By spanning the range of parameters used in previous studies, we hoped to reflect most of the possible conditions that occur in practice. It is possible that results would not be so clear if a real test was used, however, since tests differ with respect to their measurement properties. Second, only a compensatory model was used for simulating the multidimensional test responses. A noncompensatory model could produce very different results. Third, only two dimensions were generated for the multidimensional conditions. It is anticipated that DIMTEST’s power in detecting multidimensionality would be reduced in cases where there are more than two dimensions because the procedure investigates the existence of only a second dimension, but this should be studied. Fourth, the effects of variations, such as skewness and kurtosis, on the secondary ability distribution were not investigated. The secondary $\theta$ had the same distribution as the primary $\theta$. Walker et al. (2006) found that the mean and standard deviation of the secondary ability distribution play a role in the dimensional
Figure 4. T distributions and standard normal curves for unidimensional tests
structure of test data (i.e., dimensionality may go undetected for some distributions). If this is the case, power would be reduced or increased depending on the distribution of the secondary $\theta$. Finally, the effects of ability following a nonnormal distribution were not investigated. In practice it is conceivable to have skewed ability distributions, which also might affect the power to detect multidimensionality.

Conclusions

Overall, the results were fairly consistent with previous findings that DIMTEST has Type I error rates near the nominal rate and has sufficient power to detect multidimensionality (Finch & Habing, 2007; Nandakumar, 1991; Nandakumar & Stout, 1993; Stout, 1987; Stout et al., 2001). The results of the current study deviated from these findings when the sample size became large (i.e., 4,000). It appears that using more than 1,000 responses for DIMTEST may result in an increase in Type I error rates, although this effect is lower for longer tests. Further, an increase in Type I error does not necessarily inflate the empirical rate above the nominal rate because in many cases the Type I error is conservative with smaller samples. In contrast, using more responses for ATFIND has little effect on Type I error. If the distribution of item difficulties is platykurtic, more responses may be needed to ensure low Type I error and high power. The results also support the hypotheses that power will be greater if abilities have simple structure and power will increase as interability correlation decreases. The results suggest that a 50/50 split maximizes power and keeps the Type I error rate below the nominal level unless the test is short and the sample size is large. A 75/25 split controls Type I error better for short tests and large samples.

Power in identifying multidimensionality was lowest when the sample size was small, test size was small, interability correlations were high, and item difficulties were platykurtic. It appears that sample sizes of at least 2,000 are needed to ensure good power in detecting multidimensionality across most conditions. However, this varies considerably based on interability correlations and ability structure. Finally, power is also low in many of the conditions where interability correlations were high and complex structure was evident.

Declaration of Conflicting Interests

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

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.
Note

1. Holland and Rosenbaum’s approach is based on the theorem that items will be conditionally positively associated if the items are locally independent, unidimensional, and the item characteristic curves are monotone.

References


