

Evaluating DETECT Classification Accuracy and Consistency when Data Display Complex Structure

**Mark J. Gierl
Jacqueline P. Leighton
Xuan Tan**

Centre for Research in Applied Measurement and Evaluation
University of Alberta

Paper Presented at the Annual Meeting of the
National Council on Measurement in Education (NCME)

San Francisco, California, USA
April 7-10, 2006

Abstract

DETECT is an innovative and relatively new nonparametric dimensionality assessment procedure used to identify mutually exclusive, dimensionally homogeneous clusters of items using a genetic algorithm (Zhang & Stout, 1999). Because the clusters of items are mutually exclusive, this procedure is most useful when the data display approximate simple structure. In many testing situations, however, data display a complex multidimensional structure. The purpose of the current study was to evaluate DETECT item classification accuracy and consistency when the data display different degrees of complex structure using both simulated and real data analyses. Three variables were manipulated in the simulation study: The percentage of items displaying complex structure (10%, 30%, 50%), the correlation between dimensions (0.00, 0.30, 0.60, 0.75, 0.90), and the sample size (500, 1000, 1500). The results from the simulation study reveal that DETECT can accurately and consistently cluster items according to their true underlying dimension when as many as 30% of the items display complex structure, if the correlation between dimensions is less than or equal to 0.75 and the sample size is at least 1000 examinees. If 50% of the items display complex structure, then the correlation between dimensions should be less than or equal to 0.60 and the sample size be, at least, 1000 examinees. When the correlation between dimensions is 0.90, DETECT does not work well with any complex dimensional structure or sample size. These outcomes are further illustrated in two real data analyses. Implications for practice and directions for future research are discussed.

Evaluating DETECT Classification Accuracy and Consistency when Data Display Complex Structure

Dimensionality assessment is typically used to identify distinct clusters of items that, when considered collectively, help characterize the constructs measured by a set of test items. Further, dimensionality assessment is intended to help the researcher and practitioner link substantive interpretations with statistical outcomes in order to better understand the examinee-by-item interaction. With most exploratory dimensionality analyses, statistical indices and summaries are first produced to describe the underlying dimensional structure of the data. This statistical information is then interpreted substantively so that succinct terms, such as “scientific reasoning” or “algebra problem solving”, can be used to characterize the dimensions measured by a set of test items for a specific group of examinees. Thus, dimensionality assessment provides one method for connecting complex, substantively-based, test performance with statistical modeling techniques, which are designed to quantify this performance so it can be interpreted and understood across a large sample of examinees.

DETECT, the acronym for *Dimensionality Evaluation To Enumerate Contributing Traits*, is an innovative and relatively new nonparametric dimensionality assessment procedure (Kim, 1994; Zhang & Stout, 1999). It yields different types of quantitative summaries that can be used to link substantive and statistical dimensionality analyses. For example, DETECT identifies the total number of dominant dimensions underlying student performance on a set of test items; it estimates effect sizes to describe the amount of multidimensionality in a set of test items (i.e., D_{Max} index) as well as the nature of the latent structure for these items (i.e., r_{MAX}), and; it specifies which single dimension is measured best by each test item.

To achieve these outcomes, DETECT identifies mutually exclusive, dimensionally homogeneous clusters of items using a genetic algorithm. Because the clusters of items are mutually exclusive, this procedure is most useful when *approximate simple structure* prevails in the test data (Ackerman, Gierl, & Walker, 2003; Stout, Habing, Douglas, Kim, Roussos, & Zhang, 1996; Zhang & Stout, 1999). To specify these clusters, DETECT attempts to maximize the value of the DETECT index, $D(P)$. This index quantifies the degree of multidimensionality present in P .

The DETECT index is created by computing all item covariances after conditioning on the examinees' scores using the remaining items. That is,

$$D(P) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq N} \delta_{ij} E[\text{Cov}(X_i, X_j | \Theta_{TT} = \theta)],$$

where n is the number of dichotomous items on a test, P denotes the partitioning of n items into k clusters, Θ_{TT} is the test composite, X_i and X_j are scores on items i and j , and

$$\delta_{ij} = \begin{cases} 1 & \text{if items } i \text{ and } j \text{ are in} \\ & \text{the same cluster of } P, \\ -1 & \text{otherwise.} \end{cases}$$

Although many different partitions can exist in a set of test data, P^* serves as the partition that maximizes $D(P)$ [herein denoted as $D(P^*)$, but also called D_{Max} in the literature]. For instance, when the data are unidimensional, clusters of items will be found that are not homogeneous. In this case, the within-cluster conditional covariance will be positive for some pairs of items and negative for other pairs of items resulting in a $D(P^*)$ index that is close to zero. If, however, the data are multidimensional, then clusters of items will be found that have positive within-cluster conditional covariance and negative between-cluster conditional covariance, resulting in a $D(P^*)$ index that is greater than zero.

Another index that is often reported with $D(P^*)$ is r_{MAX} . To determine if the partition P^* , which produced $D(P^*)$ is, in fact, the correct partition to produce a simple structure solution, the following ratio can be computed

$$r_{Max} = \frac{D(P^*)}{\tilde{D}(P^*)},$$

where

$$\tilde{D}(P^*) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq N} | E[\text{Cov}(X_i, X_j | \Theta_{TT} = \theta)] |.$$

In other words, r_{MAX} is an index representing how well the data approximate simple structure by comparing the maximum value of the partition to the average of the absolute values for the conditional covariances across all item combinations.

DETECT was evaluated by Zhang and Stout (1999) in a series of real and simulated data studies. For the real data study, items from the analytic reasoning sections of the Graduate Record Examination and the reading comprehension section of the Law School Admissions Test were analyzed. With both datasets, DETECT produced a meaningful item partition that corresponded to the data structure of each test (i.e., content-based passages). For both analyses, the r_{MAX} index was also greater than 0.90 indicating that only real data which *approximated simple structure* were assessed [Kim (1994) claimed that r_{MAX} is ≥ 0.80 when the data display approximate simple structure; complete guidelines for interpreting $D(P^*)$ and r_{MAX} are presented later in the Results section].

For the multidimensional simulated data study, the compensatory MIRT model was used to generate data where the number of dimensions (2, 3, or 4), the number of examinees (400 or 800), and the number of items (20 or 40) were manipulated. The dependent variable was the number of replications, over 100 simulations, to which DETECT could find the correct partition P^* . The results were impressive—the number of correct runs ranged from 95 to 100 across the 24 study conditions. However, as with the real data analyses, the simulated conditions were limited to data that displayed *approximate simple structure*, as the r_{MAX} index was greater than or equal to 0.8940 across the study conditions (see Zhang & Stout, 1999, Table 6, Column 8).

Results from the Zhang and Stout (1999) study demonstrate that DETECT can identify mutually exclusive dimensionally homogeneous item clusters in diverse testing conditions when the data possess simple or approximate simple structure. In many testing situations, however, data do *not* display simple structure. Rather, the data from many educational and psychological tests display a complex multidimensional structure, meaning the items on these tests measure two or more dimensions, but to differing degrees, in the multidimensional space. The published studies conducted using DETECT, to-date, have focused on multidimensional data that display

simple or approximate simple structure. No studies, to our knowledge, have been conducted to evaluate the performance of DETECT when the data display different degrees of complex structure. Yet when complex structure prevails, as is often the case with real data analyses, the DETECT partition may be relatively poor, producing inaccurate item clusters. This point was alluded to when Zhang and Stout (1999) claimed:

It is very important to note that DETECT is still informative when approximate simple structure fails to hold. In particular, it can still locate relatively dimensionally homogeneous clusters; however, there is no longer a unique 'best' or 'correct' partition to be found by DETECT because there will be little to no separation between some of the clusters found. (p. 215)

Further, this lack of separation which results in a non-unique partition implies that the clusters produced across samples may be highly unreliable and, therefore, difficult to interpret because the clusters are not replicable across samples. This outcome is problematic because unreliable statistical results that are difficult to interpret will not lead to an increased understanding of the multidimensional structure underlying the test data. Thus, the purpose of the current study is to evaluate DETECT item classification accuracy and consistency when the data display different degrees of complex structure in a two-dimensional space. More specifically, we will determine whether DETECT is still informative when approximate simple structure fails to hold and we will identify the factors that may affect the DETECT classification accuracy and consistency when the data possess complex structure.

Method

Data Source

Examinee response vectors to a 40-item test were simulated for this study. The examinee ability estimates were generated from a bivariate normal distribution with a mean of zero and a standard deviation of one. To generate the data, realistic item parameters were selected so our results would have some generalizability to different testing situations. The item parameters contained a range of a_1 -, a_2 -, and d -parameters, typical of what can be found on multidimensional tests like the LSAT (cf. Douglas, Kim, Roussos, Stout, & Zhang, 1999). The

items were also generated so they would measure both approximate simple and complex structure.

For the items that approximate simple structure, the a_1 -parameters ranged from 0.50 to 1.10, in increments of 0.20 for θ_1 , and from 0.05 to 0.20 in increments of 0.05 for θ_2 . The same values were used for the a_2 -parameters except that the parameters ranged from 0.50 to 1.10, in increments of 0.20 for θ_2 , and from 0.05 to 0.20 in increments of 0.05 for θ_1 . The d -parameter ranged from -1.00 to 1.00 in increments of 0.50. Typically, items that lie within twenty degrees of the x- or y-axis (which correspond to the θ_1 or the θ_2 dimensions, respectively) are considered to display *approximate* simple structure (cf. Froelich, 2000; Froelich & Habing, 2001; Stout, Froelich, & Gao, 2001). For the approximate simple structure condition in our study, the first 20 items measured θ_1 . The angular directions for these items ranged from 5.71° to 10.30° . The second 20 items measured θ_2 . The angular direction for these items ranged from 79.66° to 84.26° . The item parameters and angular directions for the 40-item test that approximate simple structure are presented in Table 1.

For the items that measure complex structure, both the a_1 - and a_2 -parameters ranged from 0.50 to 1.10, in increments of 0.20, with the constraint that each pair would share a difference of only one increment. In other words, if the a_1 -parameter for an item was 0.70, then the a_2 -parameter would be 0.90. As with the approximate simple structure items, the d -parameter ranged from -1.00 to 1.00 in increments of 0.50. Data are considered to display complex structure when items measure both the first and second dimensions (i.e., a $\theta_1\theta_2$ composite). Items that lie within ten degrees of the $\theta_1\theta_2$ composite of 45° can be considered to measure a complex dimensional structure. An approximate simple structure condition and three complex structure conditions were simulated because, as previously described, the number of items measuring each dimension was a variable in this study. The first complex structure condition, referred to as complex 10%, contained the same item parameters as the simple structure

condition, except two items measuring θ_1 and two items measuring θ_2 were replaced with four items measuring the $\theta_1\theta_2$ composite. Because of this item replacement (i.e., 4 of the 40 items), 10% of the items on this 40-item test in this condition measured a complex dimensional structure. The second complex structure condition, complex 30%, included the same item parameters as the simple structure condition, except six items measuring θ_1 and six items measuring θ_2 were replaced with 12 items measuring a complex dimensional structure. This change meant that 30% of the items on the 40-item test measured the $\theta_1\theta_2$ composite. The fourth complex structure condition, complex 50%, included the same item parameters as the simple structure condition, except 10 items measuring θ_1 and 10 items measuring θ_2 were replaced with 20 items measuring a complex dimensional structure. This adjustment meant that 50% of the items on the 40-item test measured the $\theta_1\theta_2$ composite. A summary of the item parameters across all complex structure conditions is shown in Table 2.

The item parameters across all simulated conditions, as presented in Tables 1 and 2, were designed to be as similar as possible. The descriptive statistics for each condition are presented at the bottom of each table. The results reveal that the means and standard deviations for the a_1 , a_2 , and d parameters are comparable to one another, as there is little variation among the conditions.

Research Design

The purpose of this study is to evaluate the accuracy and consistency of DETECT when it is used to classify items that display complex structure in a two-dimensional space. As a result, items measuring different degrees of simple and complex structure were specified and data were simulated according to these item characteristics. Two additional factors were expected to influence item classification accuracy and consistency: The correlation between dimensions and the sample size. These two factors were also manipulated in the current study. Often, the correlation between dimensions varies on educational and psychological tests. Moreover, as the correlation between dimensions increases, the underlying data structure resembles a

unidimensional scale, possibly decreasing item classification accuracy and consistency. Because of these two considerations, the correlation between dimensions was manipulated to range from no correlation to a high correlation (i.e., 0.00, 0.30, 0.60, 0.75 or 0.90). Initially, we evaluated a more symmetric range of correlations—0.00 to 0.90 in increments of 0.30. However, this range may be overly restricted because many social science constructs have a moderate to high correlation (see, for example, the research summarized in Anastasi & Urbina, 1996; Sattler, 2001). Therefore, the correlation of 0.75 was added to ensure we would evaluate situations found commonly in educational and psychological research. Sample size was also considered an important variable because it is well known that sample size affects estimation accuracy. Samples were manipulated to range from small to large (i.e., 500, 1000, and 1500). Each condition was replicated 100 times. All analyses were conducted using the computer program DETECT where the number of dimensions was specified as 2.¹

Data Analyses

Two dependent variables were studied. The first dependent variable measured *classification accuracy* by comparing item classification between the simulated data and true item parameters. In other words, we compared the dimension each item was assigned to by DETECT with the actual dimension the item was specified to measure. The mean and standard deviation for each correct match over 100 replications was computed to evaluate classification accuracy.

The second dependent variable measured *classification consistency* by comparing the item classification between two randomly equivalent samples. That is, two randomly equivalent samples were generated for each condition, and the item classification rates across the two samples were compared. This condition was included because researchers and practitioners rarely know the dimensional structure of real data. Instead, they attempt to infer the structure, often by cross-validating their results, using a second random sample to evaluate item classification consistency between two samples. By including both analyses, we could evaluate not only how accurate DETECT classified items relative to the truth (i.e., our first analysis) but also how consistent DETECT classified items over replications (i.e., our second analysis). The

¹ Two dimensions were simulated and specified in DETECT because the number of dimensions was not a variable in this study.

mean and standard deviation for each match over 100 replications were computed to evaluate classification consistency. Each outcome in the second analysis was also compared to the actual dimension the item was specified to measure to evaluate the classification error rate in the cross-validation study (i.e., C-V error). The mean for each match between the primary and cross-validation samples relative to the true dimension the item was intended to measure over 100 replications was computed to assess cross-validation error.

Results

The results are presented in two sections. First, the item classification accuracy results are described. Recall, in these analyses the DETECT item classification rates for the simulated data were compared against the true item parameters for each dimension. Second, the item classification consistency results are presented. In these analyses, the DETECT classification rates were compared across two random samples.

Item Classification Accuracy

The mean and standard deviation for the D_{Max} and r_{Max} indices in each simulated condition are presented in Table 3. The descriptive statistics for these two indices provide a good indication of the dimensional structure for each condition according to guidelines provided in the DETECT literature. Kim (1994), for instance, suggested that a D_{Max} index less than 0.10 indicates unidimensionality; an index between 0.10 and 0.50 shows a weak amount of multidimensionality; an index between 0.51 and 1.00 can be considered a moderate amount of multidimensionality; and an index greater than 1.00 reveals a strong amount of multidimensionality. Kim also claimed that values of r_{Max} greater than or equal to 0.80 suggest the data display approximate simple structure whereas values of r_{Max} less than 0.80 suggest the data display complex structure. More recently, Roussos and Ozbek (2005) proposed more liberal D_{Max} guidelines based on their research with real and simulated data. They claimed that a D_{Max} value below 0.20 indicates very weak multidimensionality or “approximate” unidimensionality; an index between 0.20 and 0.40 indicates weak to moderate multidimensionality; an index between

0.41 and 1.0 indicates moderate to large multidimensionality; and an index greater than 1.0 indicates strong multidimensionality. Roussos and Ozbek offered no alternative r_{Max} guidelines.

Based on these guidelines, diverse dimensionality conditions were simulated in our study, as shown in Table 3. For example, the simple structure conditions contained multidimensional data that could be classified as moderate to large across all sample size conditions and most correlational conditions; only with a correlation of 0.90 was the multidimensionality in the simple structure condition deemed to be weak and to measure complex structure. For the complex structure conditions, the outcomes were much more diverse, as one would expect. Across complex 10%, for example, the conditions contained multidimensional data that could be classified as moderate to large when the correlations were 0.00, 0.30, and 0.60 with the overall structure classified as simple. When the correlation increased to 0.75, multidimensionality could be classified as moderate with complex structure. But when the correlation was 0.90, multidimensionality in complex 10% could be classified as weak with complex structure. Across complex 30%, the conditions contained multidimensional data that could be classified as moderate to large when the correlations were 0.00 and 0.30 with the overall structure classified as simple. But when the correlation was 0.60, 0.75, or 0.90, the multidimensionality could be classified as weak to moderate with the data displaying complex structure. Finally, across complex 50%, the conditions contained multidimensional data that could be classified as moderate to large when the correlation was 0.00 with the overall structure classified as simple. But when the correlation was 0.30, 0.60, 0.75, or 0.90, the multidimensionality could be classified as weak to moderate with the data displaying varying degrees of complex structure (i.e., r_{Max} had a range of values). These diverse conditions are important because they represent different degrees of multidimensionality as well as simple and complex structure, as one might expect with real data.

The item classification accuracy results are presented in Tables 4 and 5. Each table contains the item classification rates across the four dimensional structures as a function of the correlation between dimensions and sample size. The overall results are presented first (Table 4), followed

by a description of the results for Dimension 1 and 2 separately, and then for the complex structure items (Table 5). We interpret the classification rates as acceptable when the agreement between the DETECT classification and the true dimension *meet or exceed 0.90 of 90%*. This requirement is very stringent. However, it is justifiable. If the goal of the dimensionality assessment is to interpret the latent structure, then the items must be classified accurately according to the true underlying dimensional structure. In fact, these interpretations will only have value if the items are placed correctly with the dimensions they actually measure. Some may argue that this standard is too high claiming that 80% accuracy is more appropriate and feasible (80% is the rule-of-thumb that is popular in power analyses) while others may argue it is too low claiming that classification should be 100% accurate. We have selected a relatively high standard between these two alternatives, and we interpret the results relative to the 90% standard. But we also present the complete empirical results in our manuscript so other standards can be applied and other interpretations made from our analyses.

Overall Item Classification Accuracy. The overall item classification accuracy rates are presented in Table 4. These rates include the correct classifications for both simple and complex structure items. For each dimensional structure evaluated, the accuracy rates met or exceeded the 90% standard when the correlation was 0.00, 0.30, and 0.60 across all sample sizes. The only exception was found with complex 50% with a correlation of 0.60 and a sample size of 500 examinees. In this condition, accuracy was 84%. The standard deviations in these conditions were small (i.e., 0.00 to 0.07) indicating there was little variability across replications. When the correlation was increased to 0.75, the accuracy rates fell below 90% for complex 30% with a sample size of 500 examinees and all complex 50% conditions. The standard deviations in these conditions were small to moderate (i.e., 0.01 to 0.11). When the correlation was 0.90, the accuracy rates fell below 90% except for the simple structure condition with 1500 examinees—in this condition, accuracy was 91%. The standard deviations across the 0.90 correlation conditions were moderate (i.e., 0.09 to 0.14).

Classification Accuracy by Dimension. Item classification accuracy for Dimensions 1 and 2 was evaluated to identify whether errors were more prevalent on one of the two dimensions. The

outcomes were very similar to the overall item classification rates and, therefore, the results are not presented, but only described.² For each structure evaluated, the accuracy rates met or exceeded 90% for both dimensions when the correlation was 0.00, 0.30, and 0.60 across all sample sizes. The only exception was found with complex 50% with a correlation of 0.60 and a sample size of 500 examinees. In this condition, accuracy for the first and second dimensions was 84% and 83%, respectively. When the sample size was increased to 1000 examinees, accuracy on the second dimension was 88%. The standard deviations in these conditions were relatively small (i.e., 0.00 to 0.09). When the correlation was increased to 0.75, the accuracy rates fell below 90% for complex 30% with a sample size of 500 examinees (the accuracy rates were 86% and 85% for Dimensions 1 and 2, respectively) and for all complex 50% conditions (the accuracy rates ranged from 70% to 87% across the three sample size conditions). The standard deviations for the 0.75 conditions ranged from small to moderate (i.e., 0.01 to 0.13). When the correlation was 0.90, the accuracy rates fell below 90%, except for the simple structure condition with 1500 examinees (accuracy rates were 92% and 91% for Dimensions 1 and 2, respectively). The standard deviations across the 0.90 correlation conditions were moderate to large (i.e., 0.10 to 0.19).

Classification Accuracy for Complex Structure Items. Item classification accuracy was also evaluated for the complex structure items, meaning those items with an angular direction between 35° and 55° . The results are presented in Table 5. This analysis was conducted to evaluate the classification accuracy for the complex structure items, which would likely be the most difficult items to classify correctly because they measured the $\theta_1\theta_2$ ability composite. For each complex structure evaluated, the accuracy rates met or exceeded 90% only when the correlation was 0.00 for all sample size conditions. The standard deviations ranged from small to moderate (i.e., 0.02 to 0.14) indicating the classification rates varied across replications. For the remaining conditions, the effect of correlation between ability and sample size is noteworthy. With a correlation of 0.30 and a sample size of either 1000 or 1500, accuracy exceeded 90% and the standard deviations were relatively small (i.e., 0.04 to 0.09). But when the sample size

² The complete set of results are available from the first author, by request.

dropped to 500, accuracy rates fell below 90% and standard deviations increased (i.e., 0.09 to 0.16). With a correlation of 0.60 and a sample size of 1500, accuracy exceeded 90% but standard deviations were moderate to large (i.e., 0.09 to 0.16). But when the sample size was reduced to either 500 or 1000, accuracy rates fell below 90% with standard deviations remaining moderate to large (i.e., 0.10 to 0.20). With a correlation of 0.75 or 0.90, accuracy rates were below 90% for all sample sizes and standard deviations were moderate to large (i.e., 0.10 to 0.23).

Item Classification Consistency Using Cross-Validation Samples

DETECT item classification consistency was also evaluated by comparing the primary sample used in the previous analyses (see Table 3 summary) with a cross-validation sample (see Table 6 summary). As with the primary samples, the mean and standard deviation for the D_{Max} and r_{Max} indices in each simulated condition for the cross-validation samples represent diverse testing conditions with different degrees of multidimensionality as well as simple and complex structure.

Recall, classification consistency was evaluated by comparing item classification between two randomly equivalent samples because in actual analyses the dimensional structure is not often known. Rather, researchers and practitioners can infer the structure by specifying the dimensionality in the primary sample and then cross-validating their results using a second random sample where the consistency between the two analyses is used to validate the structure first specified. Each table contains the item classification rates across the four dimensional structure conditions as a function of the correlation between dimensions and sample size. The overall results are presented first (Table 7), followed by a description of the results for Dimension 1 and 2 separately, and then for the complex structure items (Table 8).

The classification consistency results, denoted as “C” in Tables 7 and 8, represent how consistently DETECT identified the same dimension in the primary and cross-validation samples. As in the previous analyses, we interpret the classification consistency rates as acceptable when the agreement between the two samples met or exceeded 0.90 or 90%. Because C can denote the consistency of both correct and incorrect classifications, the cross-validation error rate, denoted as “C-V E” was also calculated. C-V E represents how precisely DETECT identified the

same dimensions in the primary and cross-validation samples relative to the true dimension measured by each item. In other words, C-V E is the difference between C and T, where T is the actual item classification consistency rate when the true item dimension is considered. We interpret the classification error rates as acceptable when C-V E is less than or equal to 0.05 or 5%, given that the conventional Type I error rate is 5%.

Overall Item Classification Consistency and Cross-Validation Error Rate. The overall item classification consistency rates are presented in Table 7 under the column heading "C". For each dimensional structure evaluated, the consistency rates met or exceeded 90% when the correlation was 0.00, 0.30, and 0.60 across all sample size conditions. Only four exceptions were found. Classification consistency was 88% for complex 30% with a correlation of 0.60 and a sample size of 500; 86% for complex 50% with a correlation of 0.30 and a sample size of 500; and 78% and 85% for complex 50% with a correlation of 0.60 and sample sizes of 500 and 1000, respectively. The standard deviations in all conditions were small (i.e., 0.01 to 0.07). The error rate for the overall item cross-validation classification consistency results are also presented in Table 7 under the column heading "C-V E". For each dimensional structure evaluated, the error rate was less than or equal to 5% when the correlation was 0.00, 0.30, and 0.60 across all sample size conditions. In other words, the DETECT cross-validation classification results are accurate for these conditions.

When the correlation was 0.75, the consistency rates exceeded 90% for the simple structure condition across all sample sizes, but fell below 90% for both complex 10% and 30% with a sample size of 500 and for all complex 50% conditions. The standard deviations ranged from small to moderate (i.e., 0.02 to 0.11). The cross-validation error rates exceeded 5% for complex 50% when the sample size was either 500 or 1000.

When the correlation was 0.90, the consistency rates fell below 90%, ranging from 53% to 88%, for all conditions. The standard deviations ranged from small to large (i.e., 0.00 to 0.19). Further, the error rates for the 0.90 correlation conditions exceeded 5% for the simple and complex 10% conditions with 500 examinees and for both complex 30% and 50% across all

sample sizes indicating that the cross-validation results in these conditions were not only inconsistent (i.e., consistency rate below 90%) but also inaccurate (i.e., error rate above 5%).

Classification Consistency and Cross-Validation Error Rate by Dimension. Item classification accuracy for Dimensions 1 and 2 was evaluated to identify whether errors were more prevalent on one of the two dimensions. As in the previous section, the outcomes were very similar to the overall item classification rates so the results are only described. For each structure evaluated, the consistency rates met or exceeded 90% for both dimensions when the correlation was 0.00 and 0.30 across most sample size conditions (the only exception occurred when the sample size was 500 with a correlation of 0.30 for complex 50% where the accuracy rates for Dimensions 1 and 2 were 85% and 87%, respectively). The standard deviations for these conditions were also small (i.e., 0.01 to 0.08). When the correlation was 0.60, the consistency rates fell below 90% for both dimensions in complex 30% with 500 examinees (88% for both dimensions), for both dimensions in complex 50% with 500 and 1000 examinees (79% and 78% on Dimension 1 and 87% and 83% on Dimension 2, respectively), and for Dimension 2 in complex 50% with 1500 examinees (89%). The standard deviations in these conditions remained relatively small (i.e., 0.00 to 0.09). Error rates for the cross-validation consistency results were acceptable for all conditions with a correlation of 0.00, 0.30, and 0.60, except for the complex 50% condition with a correlation of 0.60 and a sample size of 500. In this condition, the error rate was 6% for both dimensions.

When the correlation was 0.75, the consistency rates exceeded 90% for the simple structure condition across all sample sizes, but fell below 90% for both complex 10% and 30% with a sample size of 500 (89% and 89% for complex 10%, and 80% and 81% for complex 30% on Dimensions 1 and 2, respectively). For complex 50%, the consistency rates were below 90% for all sample sizes (the rates ranged from 61% to 83%). The standard deviations ranged from small to moderate (i.e., 0.02 to 0.13). The cross-validation error rates exceeded 5% for at least one dimension in the complex 50% condition when the sample size was either 500 or 1000.

When the correlation was 0.90, the consistency rates fell below 90% in all conditions, ranging from 51% to 88%. The standard deviations were moderate to large (i.e., 0.12 to 0.22). The

cross-validation error rates also exceeded 5% for some, but not all, of the 0.90 correlation conditions. The C-V E rates were acceptable for the simple and complex 10% conditions with 1000 and 1500 examinees and for Dimension 1 on complex 30% with 1500 examinees. C-V E exceeded 5% for all other conditions.

Classification Consistency and Cross-Validation Error Rate for Complex Structure

Items. Item classification consistency and error rates were also evaluated for the complex structure items. The results are presented in Table 8. The consistency rates met 90% for some but not all of the conditions across the four dimensional structure conditions. For example, when the correlation was 0.00, consistency rates exceeded 90% for all complex structure conditions, but only when the sample size was 1000 or 1500. The standard deviations for these conditions ranged from small to large (i.e., 0.04 to 0.17). When the correlation was 0.30, consistency rates exceeded 90% for complex 10% and 30% with 1000 and 1500 examinees and for complex 50% with 1500 examinees. The standard deviations, again, ranged from small to large (i.e., 0.05 to 0.22). The error rates indicate the cross-validation results were accurate for all conditions with a correlation of 0.00 and 0.30.

When the correlation was 0.60, 0.75, or 0.90, consistency rates were below 90% for all conditions and the standard deviations were moderate to large (i.e., 0.10 to 0.29). The error rates, on the other hand, were acceptable for all conditions with a correlation of 0.60, except for complex 30% and 50% with 500 examinees. Alternatively, with a correlation of 0.75, the error rates were only acceptable for complex 10% with 1000 or 1500 examinees and for complex 30% with 1500 examinees. The error rates were unacceptable for all conditions with a correlation of 0.90.

Illustrative Examples

Two real data analyses were also conducted to supplement the findings presented in the simulation study. These analyses are intended to illustrate how different degrees of simple and complex structure in real data affect DETECT classification accuracy and consistency. In the first analysis, a sample of the student response data and items from the 2003 Field Trial for the SAT were used (Liu, Feigenbaum, & Walker, 2004). The SAT is a standardized test designed to

measure college readiness. Response data from a sample of 2442 students who wrote the Mathematics and Critical Reading sections were used. The Mathematics dimension was based on a subset of 36 items from the field trial. For these items, students are expected to solve unfamiliar problems using key mathematical concepts in the areas of Number and Operations; Algebra I, II, and Functions; Geometry; and Statistics, Probability, and Data Analysis. Multiple-choice and constructed-response item formats are used, but the items for both formats are scored dichotomously. The Critical Reading dimension was based on a subset of 43 items from the field trial.³ For these items, students are expected to draw inferences from text, synthesize information, distinguish between main and supporting ideas, understand word meaning, follow the logic of an argument, and recognize genres. Students solve sentence completion items in addition to critical reading items associated with short and long reading passages using content drawn from natural sciences, social studies, literary fiction, and humanities. All items are multiple choice and, therefore, scored dichotomously.

Two content-based dimensions were specified in the analyses, Mathematics and Critical Reading. The vector plot, based on the item parameters estimated from NOHARM using an exploratory two-dimensional compensatory MIRT model, is shown in Figure 1. Items from these two test sections yield an interpretable two-dimensional structure as the Mathematics items measure θ_1 whereas the Critical Reading items provide a better measure of θ_2 , meaning these items display approximate simple structure. The correlation between dimensions was moderate to strong at 0.7280. The two-dimensional model provides adequate fit to the data, as Tanaka's (1993) unweighted least squares goodness-of-fit index is 0.9734 and the root mean square residual (RMSR) is 0.0050. Further, when the dimensional structure of the items associated with the Mathematics and Critical Reading sections was evaluated using the computer program DIMTEST, with the refined bias correction method (Froelich, 2000; Froelich & Habing, 2001; Stout, Froelich, & Gao, 2001) for the 2442 student sample, two dimensionally distinct item sets

³ It is important to emphasize that we only used a subset of items from the 2003 Field Trial for the SAT. The operational version of the SAT contains more items in Mathematics and Critical Reading.

were found, $T = 11.36$, $p < 0.01$. This finding further demonstrates that the items on each section tapped distinct content-based dimensions.

The DETECT analyses for the SAT items produced a D_{Max} index of 0.4178. According to the Roussos and Ozbek (2005) guidelines, the data possess a moderate amount of multidimensionality. The r_{Max} index was 0.8192 suggesting the data display approximate simple structure (Kim, 1994). Given the results from the simulation study, we expect that DETECT would classify items consistently for Dimensions 1 and 2 because the correlation between dimensions is moderate to high ($r = 0.7280$), the data display approximate simple structure ($r_{Max} = 0.8192$), and few items measure the $\theta_1\theta_2$ composite (i.e., display complex structure). To evaluate classification consistency, two samples were produced—a primary and a cross-validation sample—by randomly splitting the original data. The 36 Mathematics items were expected to measure Dimension 1 whereas the 43 Critical Reading items were expected to measure Dimension 2. As expected, the classification results were both consistent and accurate between samples. The classification consistency for Mathematics was high at 97.2% (35/36) while the cross-validation error rate was low at 3% (1/36—in other words, only 1 of the 36 items identified in samples 1 and 2 was not actually an item that measured the mathematics dimension). The classification consistency for Critical Reading was also high at 95.4% (41/43) while the cross-validation error rate was 0%.

In the second analysis, data from the 1999 administration of the School Achievement Indicators Program (SAIP) Science Written Assessment were used (Council of Ministers of Education in Canada [CMEC], 2000). The SAIP Science Written Assessment is a standardized test designed to measure students' scientific knowledge and problem solving in both Grade 8 and Grade 11 (13- and 16-year-olds) every three to five years. The SAIP Science Written Assessment includes test items targeted to five levels of difficulty representing three broad content domains: knowledge and concepts of science, nature of science, and relationship of science to technology and societal issues. The SAIP is a dichotomously scored test containing 78 items in a multiple-choice and constructed-response format.

Two item format dimensions were specified for the SAIP Science Written Assessment, multiple choice and constructed response. Initially, response data from the 16-year-old students, who wrote one of the versions of the SAIP, were used to conduct exploratory factor analyses of the test (see Leighton, Gokiert, & Cui, 2005). Results from these initial analyses indicated that of the 78 items, only 43 items loaded reliably on any factor. Subsequent to the exploratory results, a linear factor analysis with LISREL was used to estimate the parameters for a two-dimensional model using item format as the underlying model (see Leighton et al., 2005). Of the 43 items that loaded in the previous analysis, 18 multiple-choice items were coded to measure the first dimension and 25 constructed-response items were coded to measure the second dimension. The vector plot, based on the item parameters estimated from NOHARM using an exploratory two-dimensional compensatory MIRT model with a sample of 3184 students, is shown in Figure 2. Items from these two test formats yield a two-dimensional structure, but many items measure the $\theta_1\theta_2$ composite, meaning these items possess complex structure. The correlation between dimensions is strong at 0.8019. The two-dimensional model provides reasonable fit to the data, as Tanaka's (1993) unweighted least squares goodness-of-fit index is 0.9906 and the root mean square residual (RMSR) is 0.0051. The dimensional structure of the items associated with format was also evaluated using DIMTEST, with the refined bias correction method, for the 3184 student sample. The two item formats produced dimensionally distinct item sets, $T = 3.15$, $p < 0.01$, demonstrating that the items associated with each format measure a distinct dimension.

The DETECT analyses for these items produced a D_{Max} index of 0.2094 suggesting the data possess a weak amount of multidimensionality (Kim, 1994; Roussos & Ozbek, 2005). The r_{Max} index was 0.5180 indicating the data display complex structure (Kim, 1994). Given the results from the simulation study, we expect that DETECT would not classify items consistently for Dimensions 1 and 2 because the correlation between dimensions is high ($r = 0.8019$) and the data display complex structure ($r_{Max} = 0.5180$), with many items measuring the $\theta_1\theta_2$ composite. To evaluate classification consistency, as in the previous example, two samples were produced

by randomly splitting the original data. The 25 constructed-response items were expected to measure Dimension 1 whereas the 18 multiple-choice items were expected to measure Dimension 2. The classification consistency for the constructed-response items was low at 68.0% (17/25) while the cross-validation error rate was high at 28.0% (7/25). The classification consistency for the multiple-choice items was also relatively low at 72.2% (13/18) while the cross-validation error rate was high at 11.1% (2/18).

Summary and Discussion

The purpose of this study was to evaluate DETECT item classification accuracy and consistency when the data display different degrees of complex structure in a two-dimensional space. To-date, all of the empirical studies evaluating the properties of DETECT have focused on data that display simple or approximate simple structure. In these cases, DETECT classifies items according to their underlying dimensional structure in both an accurate and a consistent manner. In many testing situations, however, the data do not display simple structure. Instead, real test data tend to display a complex multidimensional structure, meaning the items on these tests measure two or more dimensions in the multidimensional space. No research has been conducted using DETECT that focuses on multidimensional data that display different degrees of complex structure (i.e., $r_{MAX} < 0.80$), as one might find with real data. In fact, Zhang and Stout (1999) claimed that one important line of research required studies to "...investigate DETECT's capacity to find dimensionally homogeneous but non-unique clusters in the case where approximate simple structure fails" (p. 248). Our goal was to address this gap in the literature by determining whether DETECT can still classify items accurately and consistently when approximate simple structure fails to hold and, in the process, to identify the factors that may affect the DETECT classification rates when the data possess different degrees of complex structure.

To evaluate item classification accuracy, the DETECT item classification rates for simulated data with diverse characteristics were compared against the true item parameters across two dimensions. The overall classification rates were high (i.e., $> 90\%$) for all dimensional structures when the correlation was 0.00, 0.30, and 0.60 across all sample sizes. The only

exception was found with complex 50% with a correlation of 0.60 and a sample size of 500 examinees where accuracy was 84%. When the correlation was 0.75, the accuracy rates were below 90% for complex 30% with 500 examinees and for all complex 50% conditions. When the correlation was 0.90, the accuracy rates fell below 90% for almost all study conditions (the one exception occurred for the simple structure condition with 1500 examinees). When item classification accuracy was assessed separately for Dimensions 1 and 2, the results were very similar to the overall rates. Item classification accuracy was then evaluated for the complex structure items, meaning those items with an angular direction between 35° and 55° . For each structure evaluated, the accuracy rates met or exceeded 90% only when the correlation was 0.00 for all sample size conditions. With a correlation of 0.30 and a sample size of either 1000 or 1500, accuracy exceeded 90%, but when the sample size dropped to 500, accuracy rates fell below 90%. With a correlation of 0.60 and a sample size of 1500, accuracy exceeded 90% but when the sample size was reduced to either 500 or 1000, accuracy rates fell below 90%. With a correlation of 0.75 or 0.90, accuracy rates were below 90% for all sample sizes.

To evaluate item classification consistency, the DETECT item classification rates for two simulated samples, a primary and a cross-validation sample, were compared. This analysis was included because researchers and practitioners rarely know the dimensional structure of real data. Rather, they attempt to infer the structure by cross-validating their results using a second random sample. For the cross-validation conditions, the overall classification rates were high (i.e., $> 90\%$) for all dimensional structures when the correlation was 0.00, 0.30, and 0.60 across all sample sizes. The only exceptions were found in the conditions with 500 examinees (classification consistency was also below 90%, at 85%, for complex 50% with a correlation of 0.60 and sample size 1000). Further, the cross-validation error rate (i.e., C-V E) was less than or equal to 5% for these conditions. When the correlation was 0.75, the consistency rates exceeded 90% for the simple structure condition across all sample sizes, but fell below 90% for both complex 10% and 30% with a sample size of 500. For complex 50%, consistency rates were below 90% for all sample sizes. The cross-validation error rates also exceeded 5% for complex 50% when the sample size was either 500 or 1000. When the correlation was 0.90, the

consistency rates fell below 90%, ranging from 53% to 88%. Most of the classification results in these conditions were also inaccurate, as the error rates exceeded 5%. When the item classification consistency and error rates were assessed separately for Dimensions 1 and 2, the results were very similar to the overall rates. Item classification consistency was also evaluated separately for the complex structure items. When the correlation was 0.00, consistency rates exceeded 90% for all complex structure conditions, but only when the sample size was 1000 or 1500. When the correlation was 0.30, consistency rates exceeded 90% for complex 10% and 30% with 1000 and 1500 examinees and with complex 50% with 1500 examinees. The error rates indicate the cross-validation results were accurate for all conditions with a correlation of 0.00 and 0.30. When the correlation was 0.60, 0.75, or 0.90, consistency rates were below 90% for all conditions. The error rates were acceptable for all conditions with a correlation of 0.60, except for complex 30% and 50% with 500 examinees. Conversely, with a correlation of 0.75, the error rates were only acceptable for complex 10% with 1000 or 1500 examinees and for complex 30% with 1500 examinees. The error rates were unacceptable for all conditions with a correlation of 0.90.

Implications for Practice

Given the conditions evaluated in this study, we conclude that DETECT can adequately classify items—meaning the accuracy and consistency rates are equal to or greater than 90% and the cross-validation error rates are equal to or less than 5%—in a two-dimensional space for *some complex structures*. More specifically, the results from the first simulation study indicate DETECT can accurately classify items according to their true underlying dimension when as many as 30% display complex structure, when $r \leq 0.75$ and $n \geq 1000$. If 50% of the items display complex structure, then 1000 examinees or more are required and the correlation between dimensions should be ≤ 0.60 . When the correlation between dimensions is 0.90 or greater, DETECT does not work well with any dimensional structure regardless of sample size. The outcomes from the second simulation study reinforce these conclusions by highlighting the important role of sample size. The results from the cross-validation analyses suggest that DETECT can consistently and accurately classify items according to specified underlying

dimensions when as many as 30% display complex structure, when $r \leq 0.75$ and $n \geq 1000$. If 50% of the items display complex structure, then at least 1500 examinees are required and the correlation between dimensions should be ≤ 0.60 . When the correlation between dimensions is 0.90 or greater, DETECT does not work well with any dimensional structure. In short, the results from our analyses indicate that complex structure items are more difficult to identify as the correlation between dimensions increases and the sample size decreases. Therefore, researchers and practitioners are advised to use large samples ($n \geq 1500$) in their DETECT dimensionality analyses, and limit their studies to situations where the correlations between dimensions are approximately 0.60 or less when the data are expected to contain a large number of items that display complex structure (i.e., contain items that measure a composite between 35° and 55°).

Directions for Future Research

The present study provides some initial answers to the question of, how well does DETECT classify items that display a complex multidimensional structure? However, many questions remain unanswered. Therefore, additional studies should be undertaken to evaluate DETECT in other diverse and realistic testing conditions. The outcomes from these additional studies will provide researchers and practitioners with a better understanding of how DETECT performs under different testing conditions and it may yield more refined guidelines for using the D_{Max} and r_{MAX} indices.

For instance, the correlation between dimensions clearly affected the DETECT item classification results, particularly between the 0.60 and 0.90 conditions. DETECT classified items both accurately and consistently when the correlation between dimensions was 0.60 or less for most conditions in the current study. Yet, classification accuracy and consistency dropped dramatically when the correlation was 0.90. When the correlation was 0.75, complex 50% accuracy rates were consistently below the 90% standard. Taken together, these results suggest that the performance of DETECT varies when the correlation between dimensions is moderate to high. Thus, more research is needed to evaluate DETECT item classification accuracy and

consistency when correlations are specified in smaller intervals between 0.60 and 0.90 because many educational and psychological constructs are correlated in this range (see, for example, Gierl, Tan, & Wang, 2005).

Future studies can also be conducted to evaluate the effect of sample size and IRT model on the DETECT classification results. The sample sizes used in this study—500, 1000, and 1500—demonstrated that item classification improved as sample size increased. The effect of even larger sample sizes, such as 2000 and 2500, could also be evaluated for some conditions specified in the current study (e.g., conditions with a correlation of 0.90) as well as in other study conditions (e.g., correlations between 0.60 and 0.90). Also, the simulated data were generated and the real data were fit using the two-parameter compensatory MIRT model. Future studies could investigate the effects of the pseudo-guessing parameter and the use of other MIRT models (e.g., non-compensatory MIRT model) on the DETECT accuracy and consistency rates.

Perhaps the most important analyses that remain, however, are related to the performance of DETECT in higher dimensional space. Roussos and Ozbek (2005) recently noted that DETECT item classification rates were highest when two dimensions existed in the data. In the current study, our analyses were conducted by specifying the correct number of dimensions, 2, for each condition, as the number of dimensions was not a manipulated variable (see footnote 1). However, researchers and practitioners rarely know the dimensionality of their data before they conduct their analyses (i.e., they do not know that 2 should be specified as the number of dimensions when DETECT is used). Moreover, multiple dimensions (i.e., 3 or more) may be common for some educational and psychological tests. In these cases, the potential for item misclassification rises dramatically as the number of dimensions increases. Therefore, future studies should be conducted to evaluate the impact of the number of dimensions on the DETECT classification rates.

References

- Anastasi, A., & Urbina, S. (1996). *Psychological testing* (4th edition). Upper Saddle River, NJ: Prentice Hall.
- Council of Ministers of Education, Canada (2000). *Public report on science assessment: SAIP School Achievement Indicators Program 1999*. Retrieved August 12, 2002 from <http://www.cmec.ca/saip/science2/science2.en.htm>.
- Douglas, J., Kim, H., Roussos, L., Stout, W., & Zhang, J. (1999). *LSAT dimensionality analysis for the December 1991, June 1992, and October 1992*. LSAC Research Report Series. Law School Admission Council, Inc.
- Froelich, A. G. (2000). *Assessing the unidimensionality of test items and some asymptotics of parametric item response theory*. Unpublished Doctoral Dissertation. University of Illinois at Urbana-Champaign, Department of Statistics.
- Froelich, A. G., & Habing, B. (2001). *Refinements of the DIMTEST methodology for testing unidimensionality and local independence*. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.
- Gierl, M. J., Tan, X., & Wang, C. (2005, January). *Identifying cognitive dimensions that affect student performance on the New SAT—Technical Report #1: Dimensionality Results*. New York: College Examination Board.
- Leighton, J. P., Gokiert, R.J., & Cui, Y. (2005). *Using exploratory and confirmatory methods to identify the cognitive dimensions in large-scale science assessments*. Manuscript submitted for publication.
- Kim, H. R. (1994). *New techniques for the dimensionality assessment of standardized test data*. Unpublished Doctoral Dissertation. University of Illinois at Urbana-Champaign, Department of Statistics.
- Liu, J., Feigenbaum, M., Walker, M. E. (2004, April). *New SAT and New PSAT/NMSQT Spring 2003 field trial design*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

- Roussos, L. A., & Ozbek, O. (2005). *Formulation of the DETECT population parameter and evaluation of DETECT estimator bias*. Manuscript under review.
- Sattler, J. M (2001). *Assessment of children: Cognitive applications* (4th edition). San Diego, CA: Jerome M. Sattler Publisher
- Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensional assessment. *Applied Psychological Measurement, 20*, 331-354.
- Stout, W., Froelich, A. G., Gao, F. (2001). Using resampling methods to produce an improved DIMTEST procedure. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds), *Essays on Item Response Theory* (pp. 357-375). New York: Springer.
- Zhang, J. & Stout, W. (1999). The theoretical detect index of dimensionality and its application to approximate simple structure. *Psychometrika, 64*, 231-249.

Table 1

Item Parameters for the 40 Items that Display Approximate Simple Structure

Simple	Complex 10	Complex 30	Complex 50	a1	a2	d	Direction
1	X	X	X	0.50	0.05	-1.00	5.71
2				0.70	0.10	-0.50	8.13
3			X	0.90	0.15	0.00	9.46
4			X	1.10	0.20	0.50	10.30
5				0.50	0.05	1.00	5.71
6		X	X	0.70	0.10	-1.00	8.13
7				0.90	0.15	-0.50	9.46
8				1.10	0.20	0.00	10.30
9		X	X	0.50	0.05	0.50	5.71
10				0.70	0.10	1.00	8.13
11				0.90	0.15	-1.00	9.46
12		X	X	1.10	0.20	-0.50	10.30
13				0.50	0.05	0.00	5.71
14				0.70	0.10	0.50	8.13
15		X	X	0.90	0.15	1.00	9.46
16				1.10	0.20	-1.00	10.30
17			X	0.50	0.05	-0.50	5.71
18			X	0.70	0.10	0.00	8.13
19				0.90	0.15	0.50	9.46
20	X	X	X	1.10	0.20	1.00	10.30
21		X	X	0.05	0.50	-1.00	84.26
22	X	X	X	0.10	0.70	-0.50	81.84
23			X	0.15	0.90	0.00	80.51
24				0.20	1.10	0.50	79.66
25			X	0.05	0.50	1.00	84.26
26				0.10	0.70	-1.00	81.84
27		X	X	0.15	0.90	-0.50	80.51
28				0.20	1.10	0.00	79.66
29				0.05	0.50	0.50	84.26
30				0.10	0.70	1.00	81.84
31				0.15	0.90	-1.00	80.51

Table con't

32				0.20	1.10	-0.50	79.66
33			X	0.05	0.50	0.00	84.26
34		X	X	0.10	0.70	0.50	81.84
35				0.15	0.90	1.00	80.51
36			X	0.20	1.10	-1.00	79.66
37				0.05	0.50	-0.50	84.26
38				0.10	0.70	0.00	81.84
39	X	X	X	0.15	0.90	0.50	80.51
40		X	X	0.20	1.10	1.00	79.66
Simple	Mean (SD)			0.46 (0.38)	0.46 (0.38)	0.00 (0.72)	44.98 (37.09)
Complex 10	Mean (SD)			0.46 (0.38)	0.46 (0.38)	0.00 (0.71)	45.03 (37.14)
Complex 30	Mean (SD)			0.48 (0.38)	0.45 (0.38)	0.03 (0.69)	43.81 (37.22)
Complex 50	Mean (SD)			0.46 (0.38)	0.44 (0.39)	0.05 (0.72)	41.28 (37.56)

Note. X indicates the item was omitted for a particular study condition.

Table 2

Item Parameters for the 40 Items that Display Complex Structure (to be considered with Table 1)

Simple	Complex 10	Complex 30	Complex 50	a1	a2	d	Direction
X	*			0.90	0.70	-0.50	37.86
X	*			1.10	0.90	0.00	39.27
X	*			0.90	1.10	0.50	50.69
X	*			0.70	0.90	1.00	52.10
X	X			0.70	0.50	-1.00	35.52
X	X			0.50	0.70	-1.00	54.44
X	X			0.90	0.70	0.00	37.86
X	X			1.10	0.90	0.50	39.27
X	X			0.90	1.10	1.00	50.69
X	X			0.70	0.90	1.00	52.10
X	X			0.70	0.50	-0.50	35.52
X	X			0.50	0.70	-1.00	54.44
X	X	X		0.90	0.70	0.50	37.86
X	X	X		1.10	0.90	1.00	39.27
X	X	X		0.90	1.10	-1.00	50.69
X	X	X		0.70	0.90	-0.50	52.10
X	X	X		0.70	0.50	0.00	35.52
X	X	X		0.50	0.70	0.00	54.44
X	X	X		0.70	0.50	0.50	35.52
X	X	X		0.50	0.70	1.00	54.44
Simple	Mean (SD)			--	--	--	--
Complex 10	Mean (SD)			0.90 (0.16)	0.90 (0.16)	0.00 (0.79)	44.98 (7.45)
Complex 30	Mean (SD)			0.80 (0.20)	0.80 (0.20)	0.00 (0.80)	44.98 (7.93)
Complex 50	Mean (SD)			0.78 (0.20)	0.78 (0.20)	0.00 (0.73)	44.98 (8.00)

Note. X indicates the item was omitted for a particular study condition.

*The d parameters for the complex 10 items were 0.50, -0.50, -1.00, and 1.00.

Table 3

Mean and Standard Deviation for Simulated Conditions

Correlation	Sample	Structure							
		Simple		Complex 10%		Complex 30%		Complex 50%	
		D_{Max}	r_{Max}	D_{Max}	r_{Max}	D_{Max}	r_{Max}	D_{Max}	r_{Max}
0.00	500	1.65 (0.13)	0.98 (0.02)	1.44 (0.12)	0.95 (0.02)	1.04 (0.08)	0.86 (0.02)	0.65 (0.07)	0.72 (0.04)
	1000	1.63 (0.09)	0.99 (0.02)	1.41 (0.08)	0.98 (0.01)	1.04 (0.06)	0.93 (0.01)	0.64 (0.05)	0.83 (0.03)
	1500	1.62 (0.07)	1.00 (0.02)	1.39 (0.07)	0.98 (0.01)	1.02 (0.05)	0.95 (0.01)	0.63 (0.04)	0.87 (0.02)
0.30	500	1.14 (0.10)	0.94 (0.02)	0.98 (0.09)	0.88 (0.03)	0.71 (0.07)	0.77 (0.03)	0.44 (0.05)	0.60 (0.04)
	1000	1.11 (0.07)	0.98 (0.02)	0.96 (0.06)	0.95 (0.02)	0.71 (0.05)	0.87 (0.02)	0.43 (0.04)	0.72 (0.03)
	1500	1.11 (0.05)	0.99 (0.01)	0.94 (0.05)	0.96 (0.01)	0.69 (0.04)	0.90 (0.02)	0.43 (0.03)	0.78 (0.03)
0.60	500	0.63 (0.06)	0.77 (0.04)	0.55 (0.06)	0.70 (0.04)	0.41 (0.04)	0.59 (0.04)	0.26 (0.03)	0.44 (0.05)
	1000	0.62 (0.04)	0.89 (0.02)	0.54 (0.04)	0.83 (0.03)	0.40 (0.03)	0.72 (0.03)	0.25 (0.02)	0.54 (0.04)
	1500	0.62 (0.03)	0.94 (0.02)	0.52 (0.03)	0.88 (0.02)	0.39 (0.03)	0.77 (0.03)	0.24 (0.02)	0.60 (0.03)
0.75	500	0.39 (0.06)	0.56 (0.07)	0.35 (0.05)	0.52 (0.06)	0.27 (0.04)	0.44 (0.05)	0.19 (0.03)	0.34 (0.04)
	1000	0.38 (0.03)	0.72 (0.03)	0.34 (0.03)	0.67 (0.04)	0.26 (0.02)	0.57 (0.04)	0.16 (0.02)	0.41 (0.04)
	1500	0.39 (0.03)	0.81 (0.03)	0.33 (0.03)	0.74 (0.04)	0.25 (0.02)	0.62 (0.04)	0.16 (0.02)	0.46 (0.04)
0.90	500	0.19 (0.03)	0.31 (0.04)	0.19 (0.02)	0.32 (0.04)	0.17 (0.02)	0.30 (0.03)	0.16 (0.02)	0.30 (0.03)
	1000	0.16 (0.02)	0.38 (0.05)	0.15 (0.02)	0.36 (0.05)	0.13 (0.02)	0.34 (0.04)	0.11 (0.01)	0.30 (0.03)
	1500	0.15 (0.03)	0.42 (0.07)	0.13 (0.02)	0.39 (0.06)	0.12 (0.02)	0.36 (0.04)	0.09 (0.01)	0.31 (0.03)

Table 4

Overall Item Classification Accuracy

Correlation	Sample	Structure			
		Simple	Complex 10%	Complex 30%	Complex 50%
0.00	500	1.00 (0.01)	0.99 (0.01)	0.98 (0.02)	0.95 (0.03)
	1000	1.00 (0.01)	1.00 (0.01)	1.00 (0.01)	0.99 (0.01)
	1500	1.00 (0.01)	1.00 (0.00)	1.00 (0.00)	1.00 (0.01)
0.30	500	1.00 (0.00)	0.99 (0.02)	0.96 (0.03)	0.91 (0.05)
	1000	1.00 (0.01)	1.00 (0.01)	0.99 (0.02)	0.97 (0.03)
	1500	1.00 (0.00)	1.00 (0.01)	0.99 (0.01)	0.99 (0.02)
0.60	500	0.99 (0.01)	0.98 (0.02)	0.92 (0.04)	0.84 (0.07)
	1000	1.00 (0.01)	0.99 (0.01)	0.96 (0.03)	0.90 (0.05)
	1500	1.00 (0.01)	0.99 (0.02)	0.97 (0.02)	0.94 (0.04)
0.75	500	0.97 (0.05)	0.94 (0.06)	0.86 (0.07)	0.72 (0.11)
	1000	0.99 (0.01)	0.97 (0.02)	0.93 (0.04)	0.83 (0.07)
	1500	1.00 (0.01)	0.98 (0.02)	0.94 (0.03)	0.86 (0.06)
0.90	500	0.69 (0.12)	0.67 (0.13)	0.62 (0.11)	0.57 (0.09)
	1000	0.87 (0.12)	0.81 (0.14)	0.77 (0.09)	0.60 (0.10)
	1500	0.91 (0.14)	0.85 (0.14)	0.81 (0.09)	0.63 (0.10)

Table 5

Item Classification Accuracy for Complex Structure Items

Correlation	Sample	Complex Structure		
		10%	30%	50%
0.00	500	0.93 (0.14)	0.92 (0.08)	0.91 (0.06)
	1000	0.98 (0.07)	0.99 (0.03)	0.98 (0.02)
	1500	1.00 (0.02)	1.00 (0.02)	0.99 (0.02)
0.30	500	0.89 (0.16)	0.86 (0.11)	0.82 (0.09)
	1000	0.96 (0.09)	0.96 (0.06)	0.94 (0.06)
	1500	0.99 (0.05)	0.98 (0.04)	0.98 (0.04)
0.60	500	0.81 (0.20)	0.74 (0.12)	0.70 (0.11)
	1000	0.88 (0.15)	0.86 (0.10)	0.81 (0.11)
	1500	0.90 (0.16)	0.91 (0.08)	0.90 (0.09)
0.75	500	0.73 (0.23)	0.63 (0.12)	0.60 (0.13)
	1000	0.77 (0.21)	0.76 (0.12)	0.68 (0.12)
	1500	0.80 (0.19)	0.80 (0.11)	0.74 (0.10)
0.90	500	0.57 (0.23)	0.47 (0.16)	0.52 (0.10)
	1000	0.60 (0.22)	0.58 (0.13)	0.50 (0.12)
	1500	0.58 (0.19)	0.60 (0.14)	0.52 (0.11)

Table 6

Means and Standard Deviations for Cross-Validation Samples (to be compared with Table 3)

Correlation	Sample	Structure							
		Simple		Complex 10%		Complex 30%		Complex 50%	
		D_{Max}	R	D_{Max}	R	D_{Max}	R	D_{Max}	R
0.00	500	1.65 (0.12)	0.98 (0.02)	1.44 (0.11)	0.95 (0.02)	1.04 (0.09)	0.86 (0.03)	0.64 (0.06)	0.71 (0.03)
	1000	1.63 (0.08)	0.99 (0.02)	1.41 (0.07)	0.98 (0.01)	1.04 (0.06)	0.93 (0.01)	0.63 (0.04)	0.82 (0.02)
	1500	1.62 (0.07)	1.00 (0.02)	1.40 (0.06)	0.98 (0.01)	1.02 (0.04)	0.95 (0.01)	0.63 (0.03)	0.87 (0.02)
0.30	500	1.13 (0.10)	0.93 (0.03)	0.99 (0.08)	0.88 (0.03)	0.72 (0.07)	0.76 (0.04)	0.43 (0.05)	0.60 (0.04)
	1000	1.11 (0.07)	0.98 (0.02)	0.96 (0.06)	0.94 (0.02)	0.70 (0.05)	0.87 (0.02)	0.42 (0.03)	0.71 (0.03)
	1500	1.11 (0.05)	0.99 (0.02)	0.95 (0.05)	0.96 (0.01)	0.69 (0.04)	0.90 (0.02)	0.42 (0.03)	0.77 (0.03)
0.60	500	0.63 (0.07)	0.76 (0.05)	0.55 (0.06)	0.70 (0.04)	0.41 (0.04)	0.58 (0.04)	0.26 (0.03)	0.44 (0.05)
	1000	0.62 (0.04)	0.89 (0.03)	0.53 (0.04)	0.83 (0.03)	0.40 (0.03)	0.71 (0.03)	0.24 (0.02)	0.54 (0.04)
	1500	0.62 (0.04)	0.94 (0.02)	0.53 (0.03)	0.88 (0.02)	0.39 (0.02)	0.77 (0.03)	0.24 (0.02)	0.60 (0.03)
0.75	500	0.38 (0.05)	0.56 (0.06)	0.34 (0.05)	0.52 (0.05)	0.27 (0.04)	0.43 (0.05)	0.19 (0.03)	0.34 (0.04)
	1000	0.38 (0.03)	0.72 (0.04)	0.33 (0.03)	0.66 (0.04)	0.26 (0.02)	0.56 (0.04)	0.17 (0.02)	0.41 (0.04)
	1500	0.38 (0.03)	0.81 (0.03)	0.32 (0.02)	0.73 (0.03)	0.25 (0.02)	0.62 (0.03)	0.16 (0.02)	0.46 (0.04)
0.90	500	0.19 (0.02)	0.31 (0.04)	0.19 (0.03)	0.31 (0.04)	0.17 (0.02)	0.31 (0.03)	0.16 (0.01)	0.30 (0.03)
	1000	0.16 (0.03)	0.38 (0.06)	0.14 (0.02)	0.34 (0.05)	0.13 (0.02)	0.33 (0.04)	0.11 (0.01)	0.29 (0.03)
	1500	0.16 (0.02)	0.44 (0.05)	0.13 (0.02)	0.38 (0.06)	0.12 (0.02)	0.35 (0.04)	0.10 (0.01)	0.31 (0.03)

Table 7

Overall Item Classification Consistency and Cross-Validation Error Rates Between Primary and Cross-Validation Samples

Correlation n	Structure								
	Sample	Simple		Complex 10%		Complex 30%		Complex 50%	
		C	C-V E	C	C-V E	C	C-V E	C	C-V E
0.00	500	1.00 (0.01)	0.00	0.99 (0.02)	0.00	0.96 (0.03)	0.00	0.92 (0.04)	0.00
	1000	1.00 (0.01)	0.00	1.00 (0.01)	0.00	0.99 (0.01)	0.00	0.98 (0.02)	0.00
	1500	1.00 (0.01)	0.00	1.00 (0.01)	0.00	1.00 (0.01)	0.00	0.99 (0.01)	0.00
0.30	500	1.00 (0.01)	0.00	0.98 (0.02)	0.00	0.93 (0.03)	0.01	0.86 (0.05)	0.02
	1000	1.00 (0.01)	0.00	0.99 (0.01)	0.00	0.98 (0.02)	0.00	0.94 (0.04)	0.00
	1500	1.00 (0.01)	0.00	1.00 (0.01)	0.00	0.99 (0.02)	0.00	0.98 (0.02)	0.00
0.60	500	0.99 (0.01)	0.00	0.96 (0.03)	0.01	0.88 (0.05)	0.02	0.78 (0.07)	0.05
	1000	1.00 (0.01)	0.00	0.98 (0.02)	0.01	0.93 (0.03)	0.01	0.85 (0.06)	0.02
	1500	1.00 (0.01)	0.00	0.98 (0.02)	0.00	0.96 (0.03)	0.00	0.90 (0.05)	0.01
0.75	500	0.96 (0.05)	0.00	0.89 (0.07)	0.01	0.81 (0.09)	0.05	0.63 (0.11)	0.11
	1000	0.99 (0.01)	0.00	0.96 (0.02)	0.01	0.90 (0.04)	0.02	0.77 (0.08)	0.06
	1500	0.99 (0.01)	0.00	0.97 (0.02)	0.00	0.91 (0.04)	0.01	0.82 (0.06)	0.04
0.90	500	0.57 (0.12)	0.10	0.57 (0.13)	0.11	0.56 (0.10)	0.15	0.53 (0.10)	0.19
	1000	0.77 (0.14)	0.02	0.71 (0.14)	0.05	0.67 (0.11)	0.09	0.54 (0.13)	0.17
	1500	0.88 (0.15)	0.00	0.77 (0.17)	0.04	0.74 (0.12)	0.07	0.58 (0.15)	0.15

Note. C is consistency and C-V E is the cross-validation error rate.

Table 8

Item Classification Consistency and Cross-Validation Error Rates for Complex Structure Items Between Primary and Cross-Validation Samples

		Complex Structure					
		10%		30%		50%	
Correlation	Sample	C	C-V E	C	C-V E	C	C-V E
0.00	500	0.89 (0.17)	0.01	0.86 (0.11)	0.01	0.84 (0.08)	0.01
	1000	0.97 (0.09)	0.00	0.98 (0.04)	0.00	0.96 (0.04)	0.00
	1500	0.99 (0.04)	0.00	0.99 (0.03)	0.00	0.99 (0.03)	0.00
0.30	500	0.79 (0.22)	0.01	0.77 (0.11)	0.02	0.72 (0.11)	0.04
	1000	0.92 (0.14)	0.01	0.93 (0.07)	0.00	0.89 (0.08)	0.01
	1500	0.98 (0.08)	0.00	0.97 (0.06)	0.00	0.95 (0.05)	0.00
0.60	500	0.68 (0.24)	0.05	0.62 (0.15)	0.07	0.62 (0.11)	0.12
	1000	0.76 (0.23)	0.01	0.77 (0.11)	0.02	0.70 (0.12)	0.05
	1500	0.82 (0.20)	0.01	0.86 (0.11)	0.01	0.79 (0.10)	0.02
0.75	500	0.60 (0.24)	0.09	0.56 (0.15)	0.15	0.54 (0.13)	0.17
	1000	0.66 (0.24)	0.05	0.68 (0.13)	0.08	0.58 (0.13)	0.11
	1500	0.67 (0.22)	0.04	0.70 (0.14)	0.05	0.65 (0.10)	0.08

0.90	500	0.50 (0.24)	0.19	0.50 (0.16)	0.24	0.52 (0.13)	0.24
	1000	0.53 (0.29)	0.16	0.53 (0.17)	0.19	0.48 (0.16)	0.23
	1500	0.52 (0.26)	0.19	0.54 (0.19)	0.18	0.52 (0.19)	0.24

Note. C is consistency and C-V E is the cross-validation error rate.

Figure 1. The vector plot for the Mathematics and Critical Reading items on the 2003 Field Trial for the New SAT.

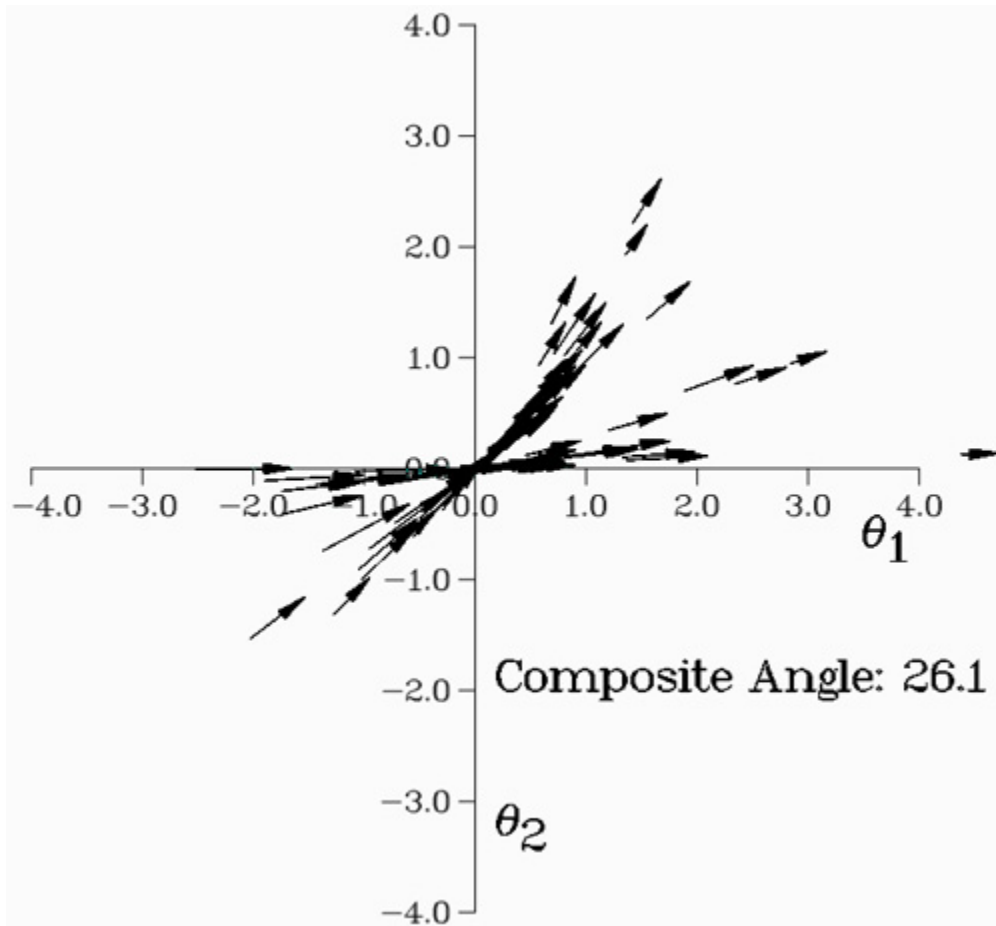


Figure 2. The vector plot for the multiple-choice and constructed-response items on the SAIP Science Written Assessment.

