PSYCHOMETRIKA—VOL. 72, NO. 1, 69–91
MARCH 2007
DOI: 10.1007/s11336-004-1257-7

CONDITIONAL COVARIANCE THEORY AND DETECT FOR POLYTOMOUS ITEMS

JINMING ZHANG

EDUCATIONAL TESTING SERVICE

This paper extends the theory of conditional covariances to polytomous items. It has been proven that under some mild conditions, commonly assumed in the analysis of response data, the conditional covariance of two items, dichotomously or polytomously scored, given an appropriately chosen composite is positive if, and only if, the two items measure similar constructs besides the composite. The theory provides a theoretical foundation for dimensionality assessment procedures based on conditional covariances or correlations, such as DETECT and DIMTEST, so that the performance of these procedures is theoretically justified when applied to response data with polytomous items. Various estimators of conditional covariances are constructed, and special attention is paid to the case of complex sampling data, such as those from the National Assessment of Educational Progress (NAEP). As such, the new version of DETECT can be applied to response data sets not only with polytomous items but also with missing values, either by design or at random. DETECT is then applied to analyze the dimensional structure of the 2002 NAEP reading samples of grades 4 and 8. The DETECT results show that the substantive test structure based on the purposes for reading is consistent with the statistical dimensional structure for either grade.

Key words: item response theory (IRT), multidimensional item response theory (MIRT), dimensionality, multidimensionality, PolyDETECT.

1. Introduction

Given a response data set, it is essential to identify its dimensional structure correctly since this is the basis of statistical analysis of the data. The simplest dimensional structure is unidimensional, which requires only one ability to explain the performance of examinees on items. Although it is the most common assumption in the analysis of response data, the unidimensionality of a set of items usually cannot be met and most tests are actually multidimensional to some extent. Many test frameworks or blueprints often stipulate that their test items measure several subscales (content strands or content areas). For instance, the current mathematics assessment of the National Assessment of Educational Progress (NAEP) measures five content strands of mathematics: *numbers and operations*, *measurement*, *geometry*, *data analysis*, and *algebra* (see Allen, Carlson, & Zelenak, 1999). In operational analysis, items are classified according to their predominant strands, such as algebra items, geometry items, and so forth, and each content-based subset of items is regarded as unidimensional. From the perspective of multidimensional item response theory (MIRT), this is equivalent to assuming that the test is multidimensional with simple structure. Although this classification according to the five content strands is commonly accepted by mathematics education experts, mathematics items can also be classified according to mathematical abilities: *conceptual understanding*, *procedural knowledge*, and *problem solving*; or according to mathematical power: *reasoning*, *connections*, and *communication* (see National Assessment Governing Board, 2002). Thus, one may obtain three different partitions of items according to content strands, mathematical abilities, or mathematical power. These partitions

This research was supported by the Educational Testing Service and the National Assessment of Educational Progress (Grant R902F980001), US Department of Education. The opinions expressed herein are solely those of the author and do not necessarily represent those of the Educational Testing Service. The author would like to thank Ting Lu, Paul Holland, Shelby Haberman, and Feng Yu for their comments and suggestions.

Requests for reprints should be sent to Jinming Zhang, Educational Testing Service, MS 02-T, Rosedale Road, Princeton, NJ 08541, USA. E-mail: jzhang@ets.org

of items into several substantively meaningful clusters are determined by test developers and subject experts. The statistical analysis of response data is usually based on such a substantive test structure as if it is the dimensional structure of the response data. However, the statistical dimensional structure results from the interaction between test items and examinees. Therefore, a substantive test structure is conceptually different from the dimensional structure. Now, the question is whether they match each other, or which substantive test structure is best in concert with the statistical dimensional structure of response data. In general, there is a great need for a procedure to identify the statistical dimensional structure of response data, specifically to identify the number of (dominant) dimensions and dimensionally homogeneous clusters of items and to verify if a target substantive test structure (approximately) matches the statistical dimensional structure. The concept of a dimensionally homogeneous cluster was rigorously defined by Zhang and Stout (1999b).

Several statistical methods are available for dimensionality analysis: multidimensional scaling, cluster analysis, and factor analysis. Multidimensional scaling is a technique for the analysis of similarity or dissimilarity among a set of objects. Given a set of similarities or distances between every pair of objects, multidimensional scaling tries to construct a configuration of the objects in a low-dimensional space such that the interpoint proximities match the original similarities or distances to the greatest extent. Oltman, Stricker, and Barrows (1990) use multidimensional scaling to analyze the test structure for the Test of English as a Foreign Language™ (TOEFL®). Cluster analysis attempts to discover natural groupings of objects (items). Grouping is done on the basis of similarities or dissimilarities (distances). Thus, a measure of similarity between objects is crucial in both cluster analysis and multidimensional scaling. Correlation coefficients or like measures of association are widely used as similarities. The purpose of factor analysis is to describe and explain the correlation among a large set of variables in terms of a small number of underlying *factors*. When directly applied to response data, factor analysis, cluster analysis, and multidimensional scaling are usually based on the item-pair covariances $\mathrm{cov}(X_{i_1}, X_{i_2})$. Typically, any two items are nonnegatively correlated in a well-designed test since examinees who earn higher scores on one item tend to earn higher scores on another. The intuitive idea of most dimensionality assessment procedures is that all items within a particular cluster are highly correlated (or have high similarities) among themselves but have relatively low correlations (or similarities) with items in a different cluster. The difficulties of grouping items into clusters are how to distinguish between high and low correlations and how to choose the number of clusters if all correlations are positive.

Many researchers use (expected) conditional covariances given an appropriately chosen subtest score to develop procedures for dimensionality assessment (Holland & Rosenbaum, 1986; Junker, 1993; Douglas, Kim, & Stout, 1994; Stout, Habing, Douglas, Kim, Roussos, & Zhang, 1996; Zhang & Stout, 1999a; Habing & Roussos, 2003). The expected conditional covariance is

$$E[\mathrm{cov}(X_{i_1}, X_{i_2} \mid Y)] = E(X_{i_1} X_{i_2}) - E[E(X_{i_1} \mid Y) E(X_{i_2} \mid Y)],$$

where Y is an observed score (e.g., the total raw score). Since $\mathrm{cov}(X_{i_1}, X_{i_2}) = E(X_{i_1} X_{i_2}) - E(X_{i_1}) E(X_{i_2})$, the expected conditional covariance and the unconditional covariance are different from each other. Instead of directly studying $E[\mathrm{cov}(X_{i_1}, X_{i_2} \mid Y)]$, Zhang and Stout (1999a) investigated the structure and properties of $E[\mathrm{cov}(X_i, X_j \mid \Theta_Y)]$ for dichotomously scored items, where Θ_Y, called the *test composite*, is an appropriately chosen composite best measured by score Y. A composite is a linear combination of the latent trait variables. Zhang and Stout (1999b) showed that the conditional covariance, given Θ_Y, will be positive if, and only if, two items measure similar constructs besides the composite Θ_Y. Hence, items can be grouped into several clusters by assigning items into the same cluster if they are positively conditionally correlated and into different clusters if they are negatively conditionally correlated. The theory of conditional covariances of dichotomous items provides a solid theoretical foundation for conditional covariance-based dimensionality assessment procedures such as DETECT (Kim, 1994; Zhang, 1996) and DIMTEST (Stout, 1987; Nandakumar & Stout, 1993). The theory also suggests that conditional covariances or conditional correlations are more appropriate and effective similarity measures than unconditional ones for use in cluster analysis and multidimensional scaling. Van Abswoude, Van der Ark, and Sijtsma (2004) did a simulation study comparing several dimensionality assessment procedures based on conditional or unconditional covariances. They found that the methods using conditional covariances were superior in finding the simulated structure to the method using unconditional covariances. However, the results of Zhang and Stout (1999a, 1999b) can only be applied to dichotomous items so far.

The purposes of this paper are to study structure and the properties of conditional covariances for polytomously scored items and to extend Zhang and Stout's (1999a, 1999b) results to tests that incorporate polytomous items. The remainder of the paper is organized as follows: Section 2 presents some theoretical results concerning the structure of (expected) conditional covariance of two items, dichotomously or polytomously scored. In Section 3 two types of sample conditional covariances are proposed for different situations. In Section 4 the theory of conditional covariances developed in Section 2 is used to theoretically justify the DETECT procedure for a test with polytomous items. Section 5 shows some simulation results using PolyDETECT, a computer program based on the procedure developed in this paper. In Section 6 PolyDETECT is applied to analyze the 2002 NAEP reading data. Finally, Section 7 summarizes the results and provides further discussion.

2. Theory of Conditional Covariances

Suppose there is a test with n items and examinees' responses to item i can be classified into $m_i + 1$ ordered categories ($m_i \geq 1$), scored $0, 1, \ldots, m_i$, respectively. Let X_i be the score on item i for a randomly selected examinee from a certain population. When $m_i = 1$, then X_i is a binary variable. MIRT assumes that the performance of an examinee on a test can be explained by a latent trait (ability) vector. The underlying latent trait vector is denoted as $\Theta = (\Theta_1, \Theta_2, \ldots, \Theta_d)'$, where Θ is a column vector and d is the number of dimensions. The kth *item category response function* (ICRF) is defined as the probability of getting score k on an item for a randomly selected examinee with ability vector $\theta = (\theta_1, \theta_2, \ldots, \theta_d)'$. That is,

$$P_{ik}(\boldsymbol{\theta}) = P(X_i = k \mid \Theta = \boldsymbol{\theta}), \qquad k = 0, 1, \ldots, m_i. \tag{1}$$

$P_{ik}(\boldsymbol{\theta})$ is also called the *item category characteristic function*. The *item response function* (IRF) is defined as the expected item score given the ability vector $\boldsymbol{\theta}$, that is,

$$\mu_i(\boldsymbol{\theta}) \equiv E[X_i \mid \Theta = \boldsymbol{\theta}] = \sum_{k=1}^{m_i} k P_{ik}(\boldsymbol{\theta}). \tag{2}$$

When the item is dichotomously scored, $\mu_i(\boldsymbol{\theta}) = P_{i1}(\boldsymbol{\theta}) = P(X_i = 1 \mid \boldsymbol{\theta})$. The IRF is assumed to be (monotone) increasing, that is, the expected score of an item increases monotonically when at least one of the abilities increases. Usually, it is also assumed that *local independence* holds, that is, $X_1, X_2, \ldots, X_n$ are independent given Θ. Some researchers (McDonald, 1994; Stout et al., 1996) suggest using a weak version of local independence,

$$\text{cov}(X_{i_1}, X_{i_2} \mid \Theta = \boldsymbol{\theta}) = 0$$

for all $\boldsymbol{\theta}$ and $1 \leq i_1 < i_2 \leq n$; that is, items are *pairwise locally uncorrelated*.

In this paper, a test is said to be *d-dimensional* if d is the minimal number of abilities required to produce a pairwise locally uncorrelated, monotone IRF model. When $d = 1$, the test is called *unidimensional*. Here, the pairwise locally uncorrelated condition is assumed instead of the local independence. Note that if the local independence is required instead in the above definition, the number of dimensions under such an alternative definition should be larger than or equal to the number of dimensions defined in this paper because the local independence is stronger than the pairwise locally uncorrelated condition. However, these two numbers of dimensions should be the same in most cases of real test data since the local independence is expected to hold in practice if items are pairwise locally uncorrelated. Further research is needed to find conditions under which the pairwise locally uncorrelated condition is equivalent to local independence or the two definitions of dimensionality are the same.

A *composite*, Θ_α, of the latent vector Θ is defined to be a linear combination of Θ, that is, $\Theta_\alpha = \alpha'\Theta = \sum_{j=1}^{d} \alpha_j \Theta_j$, where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d)'$ is any fixed constant vector and the standardized vector of $\boldsymbol{\alpha}$ is called the direction of the composite Θ_α. The elements, $\alpha_1, \ldots, \alpha_d$, are also called the weights of the composite. Theoretically, it is convenient to assume that Θ_α is standardized such that its variance is one. In practice, it is usually assumed that the summation of weights is one. Nevertheless, composites used as conditioning variables are equivalent as long as they have the same direction.

To prove the properties of (expected) conditional covariances, this paper makes two assumptions, which are either the same as, or parallel to, the corresponding assumptions for dichotomous items in Zhang and Stout (1999a).

2.1. Two Assumptions

The first assumption is that the latent trait vector Θ has a multivariate normal distribution, $\Theta \sim N(\mathbf{0}, \Sigma)$, where $\Sigma = (\rho_{ij})$ is a $d \times d$ positive definite matrix with $\rho_{ij} \geq 0$. Without loss of generality, one may assume $\rho_{jj} = 1$ for $j = 1, 2, \ldots, d$ and Σ is the correlation matrix. The second assumption is that each item is modeled by a generalized multidimensional (polytomous) compensatory model defined below.

An item is said to be modeled by a *generalized multidimensional* (*polytomous*) *compensatory model* if its IRF can be written as

$$\mu_i(\boldsymbol{\theta}) = H_i(\mathbf{a}_i'\boldsymbol{\theta}) \equiv H_i\left(\sum_{j=1}^{d} a_{ij}\theta_j\right), \tag{3}$$

where $\mathbf{a}_i = (a_{i1}, a_{i2}, \ldots, a_{id})'$, $a_{i1}, a_{i2}, \ldots, a_{id}$ are nonnegative and not all zero, and $H_i(x)$ is any nondecreasing differentiable function (i.e., $H_i'(x) \geq 0$). The $\mathbf{a}_i$ is called the *discrimination parameter vector*, and $H_i(\cdot)$ the *link function*. According to (3), the ability change in a dimension with a larger discrimination parameter has larger impact on the expected item score than the change in another dimension with a smaller discrimination parameter. This model is considered to be compensatory because through $\sum_{j=1}^{d} a_{ij}\theta_j$ high ability values on some dimensions can compensate for low values on other dimensions. This is an extension of the generalized compensatory model for a dichotomous item proposed by Zhang and Stout (1999a). As discussed there, the generalized compensatory model includes many currently used latent trait models for dichotomous items such as the multidimensional two-parameter logistic (2PL) model (Reckase, 1985; Reckase & McKinley, 1991) and the multidimensional compensatory normal ogive model. As shown later, the family of generalized compensatory models also includes the multidimensional compensatory versions of the generalized partial credit model and the graded response model.

The ICRF of a generalized partial credit model for a unidimensional case (Muraki, 1992) can be written as

$$P_{ik}(\theta) = \frac{\exp\{(k+1)a_i\theta - b_{ik}\}}{\sum_{j=0}^{m_i} \exp\{(j+1)a_i\theta - b_{ij}\}} \qquad \text{for} \qquad k = 0, 1, \ldots, m_i, \tag{4}$$

where a_i and b_{ik} are unknown item parameters, and $b_{im_i} = (m_i + 1)b_{i0}$. The ICRF of a multidimensional compensatory generalized partial credit model is defined as

$$P_{ik}(\boldsymbol{\theta}) = P(X_i = k \mid \theta) = \frac{\exp\{(k+1)\mathbf{a}_i^{'}\theta - b_{ik}\}}{\sum_{j=0}^{m} \exp\{(j+1)\mathbf{a}_i^{'}\theta - b_{ij}\}} \qquad \text{for} \quad k = 0, 1, \ldots, m_i, \qquad (5)$$

where $\mathbf{a}_i = (a_{i1}, a_{i2}, \ldots, a_{id})^{'}$ is the discrimination parameter vector. The corresponding IRF can be written as

$$\mu_i(\boldsymbol{\theta}) = H_{i1}(\mathbf{a}_i^{'}\boldsymbol{\theta}),$$

where $H_{i1}(z)$ is a link function and

$$H_{i1}(z) = \frac{\sum_{k=1}^{m_i} k \exp\{(k+1)z - b_{ik}\}}{\sum_{k=0}^{m_i} \exp\{(k+1)z - b_{ik}\}}.$$

It is not difficult to verify that $H_{i1}(z)$ is a smooth increasing function of z. This shows that the multidimensional compensatory generalized partial credit model defined in (5) is a special case of a generalized compensatory model with link function $H_{i1}(\cdot)$.

The graded response model (homogeneous 2PL case; see Samejima, 1969, 1972) is defined in the form of a reversed cumulative function,

$$P_{ik}^{*}(\theta) = \text{Prob}\{X_i \geq k \mid \theta\} = \frac{\exp[a_i\theta - d_{ik}]}{1 + \exp[a_i\theta - d_{ik}]}, \qquad k = 1, 2, \ldots, m_i,$$

where a_i and d_{ik} $(k = 1, 2, \ldots, m_i)$ are parameters. The multidimensional extension of the graded response model is defined as

$$P_{ik}^{*}(\boldsymbol{\theta}) = \text{Prob}\{X_i \geq k \mid \boldsymbol{\theta}\} = \frac{\exp[\mathbf{a}_i^{'}\boldsymbol{\theta} - d_{ik}]}{1 + \exp[\mathbf{a}_i^{'}\boldsymbol{\theta} - d_{ik}]}, \qquad k = 1, 2, \ldots, m_i.$$

The corresponding IRF can be obtained

$$\mu_i(\boldsymbol{\theta}) = \sum_{k=1}^{m_i} P_{ik}^{*}(\boldsymbol{\theta}) = H_{i2}(\mathbf{a}_i^{'}\boldsymbol{\theta}),$$

where $H_{i2}(z)$ is a link function and

$$H_{i2}(z) = \sum_{k=1}^{m_i} \frac{\exp[z - d_{ik}]}{1 + \exp[z - d_{ik}]}.$$

It is obvious that $H_{i2}(\cdot)$ is a smooth increasing function. Thus, the multidimensional graded response model defined here also belongs to the family of generalized compensatory models.

2.2. Properties of Conditional Covariances

Under the two assumptions of Section 2.1, all of the results in Zhang and Stout (1999a, 1999b) for dichotomous items still hold for polytomous items. Their proofs are also similar. For a given composite $\Theta_\alpha = \boldsymbol{\alpha}^{'}\Theta$ (i.e., fixed $\boldsymbol{\alpha}$), define

$$\lambda_{i_1 i_2}(\boldsymbol{\alpha}) = \text{cov}(\mathbf{a}_{i_1}^{'}\Theta, \mathbf{a}_{i_2}^{'}\Theta \mid \Theta_\alpha), \qquad (6)$$

where $\mathbf{a}_{i_1}$ and $\mathbf{a}_{i_2}$ are the discrimination parameter vectors of items i_1 and i_2, respectively. By Lemma 1 of Zhang and Stout (1999a), which can be derived from Theorem 2.5.1 of Anderson (1984),

$$\lambda_{i_1 i_2}(\alpha) = \mathbf{a}_{i_1}^{'} \Sigma\, \mathbf{a}_{i_2} - \frac{(\mathbf{a}_{i_1}^{'} \Sigma\, \boldsymbol{\alpha})(\mathbf{a}_{i_2}^{'} \Sigma\, \boldsymbol{\alpha})}{\boldsymbol{\alpha}^{'} \Sigma\, \boldsymbol{\alpha}}. \qquad (7)$$

The following theorem extends Theorem 1 in Zhang and Stout (1999a) to polytomously scored items.

Theorem 1. *For a given composite* $\Theta_\alpha = \boldsymbol{\alpha}'\Theta$,

$$\mathrm{Sgn}[\mathrm{Cov}(X_{i_1}, X_{i_2} \mid \Theta_\alpha)] = \mathrm{Sgn}[\lambda_{i_1 i_2}(\alpha)], \tag{8}$$

where Sgn(x) is the sign function that gives the sign (+, −, *or* 0) *of* x. *That is, for all* θ, $\lambda_{i_1 i_2}(\alpha)$ *and* $\mathrm{Cov}(X_{i_1}, X_{i_2} \mid \Theta_\alpha = \theta)$ *always have the same sign.* Moreover, $\mathrm{Cov}(X_{i_1}, X_{i_2} \mid \Theta_\alpha = \theta)$ *is a strictly increasing function of* $\lambda_{i_1 i_2}(\alpha)$ *when* $d > 2$, *and* $\lambda_{i_1 i_1}(\alpha)$ *and* $\lambda_{i_2 i_2}(\alpha)$ *are fixed. These results also hold for the expected conditional covariance* $E[\mathrm{Cov}(X_{i_1}, X_{i_2} \mid \Theta_\alpha)]$.

Proof. First the conditional covariance can be decomposed as

$$\mathrm{cov}(X_{i_1}, X_{i_2} \mid \Theta_\alpha) = \mathrm{cov}(E(X_{i_1}|\Theta), E(X_{i_2}|\Theta) \mid \Theta_\alpha) + E(\mathrm{cov}(X_{i_1}, X_{i_2}|\Theta) \mid \Theta_\alpha).$$

Then by the condition that items are pairwise locally uncorrelated, we obtain

$$\mathrm{cov}(X_{i_1}, X_{i_2} \mid \Theta_\alpha) = \mathrm{cov}[\mu_{i_1}(\Theta), \mu_{i_2}(\Theta) \mid \Theta_\alpha],$$

where $\mu_{i_1}(\boldsymbol{\theta})$ and $\mu_{i_2}(\boldsymbol{\theta})$ are the item response functions for items i_1 and i_2, respectively. When each item is modeled by a generalized compensatory model,

$$\mathrm{cov}(X_{i_1}, X_{i_2} \mid \Theta_\alpha) = \mathrm{cov}[H_{i_1}(\mathbf{a}_{i_1}'\Theta), H_{i_2}(\mathbf{a}_{i_2}'\Theta) \mid \Theta_\alpha],$$

where $H_{i_1}(\boldsymbol{\theta})$ and $H_{i_2}(\boldsymbol{\theta})$ are the link functions of items i_1 and i_2, respectively. By Lemma 3 of Zhang and Stout (1999a), (8) is obtained and $\mathrm{cov}(X_{i_1}, X_{i_2} \mid \Theta_\alpha = \theta)$ for any given θ is a strictly increasing function of $\lambda_{i_1 i_2}(\alpha)$ when $d > 2$, and $\lambda_{i_1 i_1}(\alpha)$ and $\lambda_{i_2 i_2}(\alpha)$ are fixed.

The expected conditional covariance of X_{i_1} and X_{i_2} given Θ_α is given by

$$E[\mathrm{cov}(X_{i_1}, X_{i_2} \mid \Theta_\alpha)] = \int_{-\infty}^{\infty} \mathrm{cov}(X_{i_1}, X_{i_2} \mid \Theta_\alpha = \theta) f_{\boldsymbol{\alpha}}(\theta)\, d\theta, \tag{9}$$

where $f_{\boldsymbol{\alpha}}(\theta)$ is the (normal) density function of the composite Θ_α that is determined by the composite direction $\boldsymbol{\alpha}$ and the (normal) distribution of the latent vector $\boldsymbol{\Theta}$. Notice that (8) holds for any given θ value. Therefore,

$$\mathrm{Sgn}[E[\mathrm{cov}(X_{i_1}, X_{i_2} \mid \Theta_\alpha)]] = \mathrm{Sgn}[\lambda_{i_1 i_2}(\alpha)],$$

and $E[\mathrm{cov}(X_{i_1}, X_{i_2} \mid \Theta_\alpha)]$ is a strictly increasing function of $\lambda_{i_1 i_2}(\alpha)$ when $d > 2$, and $\lambda_{i_1 i_1}(\alpha)$ and $\lambda_{i_2 i_2}(\alpha)$ are fixed. □

Theorem 1 shows that the sign of the conditional covariance of two items is exactly the same as that of the two composites with corresponding discrimination vectors as their directions. Moreover, the (expected) conditional covariance of two items has a positive monotone relationship with the conditional covariance of the two composites. When the number of dimensions is two, the same result of Corollary 2 in Zhang and Stout (1999a) can be obtained for polytomous items from Theorem 1.

Corollary 1. *If a test is two dimensional, then for any given* θ,

$$\mathrm{Cov}(X_{i_1}, X_{i_2} \mid \Theta_\alpha = \theta) \begin{cases} > 0, & \textit{if the discrimination parameter vectors of items } i_1 \textit{ and } i_2 \textit{ are on the same side of vector } \boldsymbol{\alpha}; \\ = 0, & \textit{if at least one of the discrimination parameter vectors is in the same direction as vector } \boldsymbol{\alpha}; \\ < 0, & \textit{if the discrimination parameter vectors of items } i_1 \textit{ and } i_2 \textit{ are on different sides of vector } \boldsymbol{\alpha}; \end{cases} \tag{10}$$

and $\mathrm{Sgn}[E[\mathrm{Cov}(X_{i_1}, X_{i_2} \mid \Theta_\alpha)]] = \mathrm{Sgn}[\mathrm{Cov}(X_{i_1}, X_{i_2} \mid \Theta_\alpha = \theta)]$.

Corollary 1 indicates that the sign of (expected) conditional covariance is solely determined by two discrimination parameter vectors and the composite direction when a test is two dimensional.

Let $V = \{\mathbf{a} = (a_1, a_2, \ldots, a_d)^{'}: \text{all } a_i \text{ are real numbers}\}$. The *inner product* in V is defined by $\langle \mathbf{a}_i, \mathbf{a}_j \rangle = \mathbf{a}_i^{'} \Sigma\, \mathbf{a}_j$ for any $\mathbf{a}_i, \mathbf{a}_j \in V$, where Σ is the correlation matrix of the latent trait vector Θ. Then V is a d-dimensional Euclidean space. The length of vector $\mathbf{a}$ is defined as $\|\mathbf{a}\| \equiv \sqrt{\langle \mathbf{a}, \mathbf{a} \rangle}$, and the angle β between $\mathbf{a}_i$ and $\mathbf{a}_j$ is defined as

$$\beta = \cos^{-1}\left(\frac{\langle \mathbf{a}_i, \mathbf{a}_j \rangle}{\|\mathbf{a}_i\|\,\|\mathbf{a}_j\|}\right).$$

Let

$$V_{\boldsymbol{\alpha}}^{\perp} = \{\mathbf{a} \in V : \langle \mathbf{a}, \boldsymbol{\alpha} \rangle = 0\}.$$

$V_{\boldsymbol{\alpha}}^{\perp}$ is a $(d-1)$-dimensional subspace that is orthogonal to $\boldsymbol{\alpha}$. When $d = 3$, $V_{\boldsymbol{\alpha}}^{\perp}$ is a plane perpendicular to $\boldsymbol{\alpha}$. The following theorem extends Theorem 2 of Zhang and Stout (1999a) and Lemma 3 of Zhang and Stout (1999b) to polytomously scored items.

Theorem 2. *For a given composite* $\Theta_\alpha = \boldsymbol{\alpha}^{'}\Theta$,

$$\lambda_{i_1 i_2}(\alpha) = \|\mathbf{a}_{i_1}^{\perp}\|\,\|\mathbf{a}_{i_2}^{\perp}\| \cos \beta_{i_1 i_2}, \tag{11}$$

where $\mathbf{a}_i^{\perp}$ *is the projection of the discrimination parameter vector* $\mathbf{a}_i$ *on the space* $V_{\boldsymbol{\alpha}}^{\perp}$, *and* $\beta_{i_1 i_2}$ *is the angle between* $\mathbf{a}_{i_1}^{\perp}$ *and* $\mathbf{a}_{i_2}^{\perp}$ $(0 \leq \beta_{i_1 i_2} \leq \pi)$. *Thus, if at least one of the discrimination parameter vectors* $\mathbf{a}_{i_1}$ *and* $\mathbf{a}_{i_2}$ *is in the same direction as* $\boldsymbol{\alpha}$, *then, for any* θ,

$$\mathrm{Cov}(X_{i_1}, X_{i_2} \mid \Theta_\alpha = \theta) = 0, \tag{12}$$

and

$$E[\mathrm{Cov}(X_{i_1}, X_{i_2} \mid \Theta_\alpha)] = 0. \tag{13}$$

Otherwise*,*

$$\mathrm{cov}(X_{i_1}, X_{i_2} \mid \Theta_\alpha = \theta) \begin{cases} > 0, & \textit{if } \beta_{i_1 i_2} < \pi/2 \\ = 0, & \textit{if } \beta_{i_1 i_2} = \pi/2 \\ < 0, & \textit{if } \beta_{i_1 i_2} > \pi/2 \end{cases} \tag{14}$$

and $\mathrm{Sgn}[E[\mathrm{cov}(X_{i_1}, X_{i_2} \mid \Theta_\alpha)]] = \mathrm{Sgn}[\mathrm{cov}(X_{i_1}, X_{i_2} \mid \Theta_\alpha = \theta)]$ *for any* θ. *Moreover, the magnitudes of* $\mathrm{cov}(X_{i_1}, X_{i_2} \mid \Theta_\alpha = \theta)$ *for any* θ *and* $E[\mathrm{cov}(X_{i_1}, X_{i_2} \mid \Theta_\alpha)]$ *are strictly decreasing functions of* $\beta_{i_1 i_2}$ *for* $\beta_{i_1 i_2} \in [0, \pi]$ *when* $d > 2$, *and* $\|\mathbf{a}_{i_1}^{\perp}\|$ *and* $\|\mathbf{a}_{i_2}^{\perp}\|$ *are fixed.*

Proof. Similar to the proof of Lemma 3 of Zhang and Stout (1999b), $\mathbf{a}_{i_1}$ and $\mathbf{a}_{i_2}$ can be uniquely decomposed into

$$\mathbf{a}_{i_1} = c_1 \alpha + \mathbf{a}_{i_1}^{\perp} \tag{15}$$

$$\mathbf{a}_{i_2} = c_2 \alpha + \mathbf{a}_{i_2}^{\perp} \tag{16}$$

where c_1 and c_2 are constants, and $\mathbf{a}_{i_1}^{\perp}, \mathbf{a}_{i_2}^{\perp} \in V_{\boldsymbol{\alpha}}^{\perp}$. By (7), (15) and (16), one can obtain (11). By Theorem 1 and (11), one obtains (12) and (14). Since $\lambda_{i_1 i_1}(\alpha) = \|\mathbf{a}_{i_1}^{\perp}\|^2$ and $\lambda_{i_2 i_2}(\alpha) = \|\mathbf{a}_{i_2}^{\perp}\|^2$, by Theorem 1, $\mathrm{cov}(X_{i_1}, X_{i_2} \mid \Theta_\alpha = \theta)$ is a strictly decreasing function of $\beta_{i_1 i_2}$ for $\beta_{i_1 i_2} \in [0, \pi]$ when $d > 2$, and $\|\mathbf{a}_{i_1}^{\perp}\|$ and $\|\mathbf{a}_{i_2}^{\perp}\|$ are fixed. By (9), the results for expected conditional covariance are obtained. □

When applied to practical cases, the composite used as a conditioning variable is typically the test composite. Informally, the test composite is a common composite that all items try to measure, at least partially. Thus, whether two items in a test measure similar constructs depends on what they measure besides the test composite. This leads to the concept of *dimensionally homogeneous or heterogeneous* items. As defined in Zhang and Stout (1999b), dimensionally homogeneous items means that the items are close to each other in the subspace that is perpendicular to the direction of the test composite. That is, we judge whether two items are dimensionally homogeneous or not by examining whether the projections of their discrimination vectors on that subspace are close to each other or not. The angle $\beta_{i_1 i_2}$ is the measure of the closeness or dimensional homogeneity of two items. Specifically, two items are dimensionally homogeneous if the angle is less than $\pi/2$, and dimensionally heterogeneous if the angle is larger than $\pi/2$. Theorem 2 shows that the greater the degree of dimensional homogeneity of two items (i.e., the smaller the angle $\beta_{i_1 i_2}$), the larger the positive conditional covariance is. Thus, high-dimensional homogeneity is associated with large positive conditional covariance, and high-dimensional heterogeneity is associated with negative conditional covariance with large magnitude. Hence, conditional covariances can be used to group test items into clusters such that any two items in the same cluster are positively conditionally correlated.

A test has an *approximate simple structure* if it consists of several clusters with each cluster consisting of dimensionally homogeneous items while items from different clusters are dimensionally heterogeneous. By Theorem 2 and the definitions of dimensionally homogeneous and heterogeneous items, a test has an approximate simple structure if, and only if, there exists a partition $\mathcal{P}^* = \{A_1, A_2, \ldots, A_K\}$ such that

$$\beta_{i_1 i_2} \begin{cases} < \frac{\pi}{2}, & \text{if items } i_1 \text{ and } i_2 \text{ come from the same } A_k; \\ > \frac{\pi}{2}, & \text{if items } i_1 \text{ and } i_2 \text{ come from two different } A_k\text{,s;} \end{cases} \tag{17}$$

where $\beta_{i_1 i_2}$ is the angle between $\mathbf{a}_{i_1}^{\perp}$ and $\mathbf{a}_{i_2}^{\perp}$ $(0 \leq \beta_{i_1 i_2} \leq \pi)$, $1 \leq i_1 \neq i_2 \leq n$, whereas $\mathbf{a}_{i_1}^{\perp}$ and $\mathbf{a}_{i_2}^{\perp}$ are the projections of the discrimination parameter vectors $\mathbf{a}_{i_1}$ and $\mathbf{a}_{i_2}$ on the space $V_{\boldsymbol{\alpha}}^{\perp}$, respectively. In other words, a test has an approximate simple structure if, and only if, there exists a partition such that the conditional covariances given a test composite are positive for every item pair from the same cluster, and negative for every item pair from the different clusters. This type of tests will be discussed further in Section 4.

3. Sample Expected Conditional Covariance

The previous section presented some important properties of (expected) conditional covariances given a composite that can be utilized to analyze the dimensional structure of a test. However, a composite is a latent variable, which cannot be used in practice. This section discusses how to choose appropriately a manifest variable as a conditioning variable to form sample expected conditional covariances.

3.1. A Composite Scale Score as a Conditioning Variable

In this approach, an appropriate composite score is estimated for each examinee. The idea is that a unidimensional calibration program, such as PARSCALE (Muraki & Bock, 1997), is used to produce composite scale scores for all examinees. Then examinees are partitioned into homogeneous ability groups according to their composite scale scores.

There are two ways to produce composite scale scores for the conditioning purpose: the unidimensional approximation approach and the simple structure approach. The unidimensional approximation approach treats the whole response data as unidimensional to produce ability estimates. These ability estimates are actually the estimates of a *reference* composite (see Wang, 1986). The simple structure approach regards the whole response data as multidimensional with simple structure. Since an educational test usually has several target subscales to measure, the simple structure assumption is widely used in practice, as mentioned in Section 1. Under the simple structure assumption, each item is regarded as measuring only one subscale, and each content-based subtest (items measuring the same subscale) is considered to be unidimensional and is calibrated separately using a unidimensional calibration program. Then a composite score, formed using appropriately chosen weights, can be used as a conditioning variable. Users may choose these weights to be proportional to the perfect raw scores of subtests or to be inversely proportional to the average measurement error variance on each subscale.

Examinees are then stratified into groups according to their composite scores. The percentiles of composite scores may be used as cut-points in forming groups. Generally speaking, the more the groups (cut-points), the more homogeneous examinees's composite abilities within each group are. However, when there are too many groups, the number of examinees in a group will be too small and consequently damage the accuracy of conditional covariance estimation. Thus, the number of groups and the number of examinees in each group should be chosen in a balanced way such that the composite scores of examinees in the same group are similar enough and the number of examinees in each group is large enough to achieve optimal conditional covariance estimation.

Let J_k be the number of students in the kth stratum. Denote $J = \sum_k J_k$ as the total number of students. Within Group k, calculate sample covariances of any item pairs

$$\widehat{\text{cov}}(X_{i_1}, X_{i_2} \mid \text{Group } k) = \frac{1}{J_k} \sum_{j=1}^{J_k} (x_{i_1 jk} - \overline{x}_{i_1 k})(x_{i_2 jk} - \overline{x}_{i_2 k}),$$

where x_{ijk} is the score on item i for the jth examinee in the kth stratum (Group k), and $\overline{x}_{ik} = (1/J_k) \sum_{j=1}^{J_k} x_{ijk}$. Then, one may construct an estimator of $E[\text{cov}(X_{i_1}, X_{i_2} \mid \Theta_\alpha)]$,

$$\widehat{E\text{cov}}_{i_1 i_2}(\alpha) = \sum_k \frac{J_k}{J} \widehat{\text{cov}}(X_{i_1}, X_{i_2} \mid \text{Group } k), \tag{18}$$

where the α are the weights that are used in forming the composite in the simple structure approach or the weights of the reference composite in the unidimensional approximation approach. The estimator of expected conditional correlation is

$$\widehat{E\text{Cor}}_{i_1 i_2}(\alpha) = \frac{\widehat{E\text{cov}}_{i_1 i_2}(\alpha)}{\sqrt{\widehat{E\text{cov}}_{i_1 i_1}(\alpha)\widehat{E\text{cov}}_{i_2 i_2}(\alpha)}}. \tag{19}$$

3.2. An Observed Raw Score as a Conditioning Variable

In practice, observed raw scores are usually more ready to be used than composite scale scores. When the test length is long enough, the conditional covariance given an observed score can be regarded as an approximation to conditional covariance given an appropriate value of the test composite, which is best measured by the observed score in the sense that the observed score has maximum discriminating power in the direction of this composite. For details, see Zhang and Stout (1999a).

First, the total score, $T = \sum_{i=1}^{n} X_i$, can be used as a conditioning variable. Examinees are stratified into several groups according to their total scores. If the perfect (highest) total score is N (Nn), then initially there are $(N + 1)$ groups of examinees. However, some groups (typically with extremely low or high scores) may be too small to accurately estimate covariance within groups. Usually, a lower bound is selected to eliminate such groups. If the number of examinees in a group is fewer than the lower bound, then this group merges to its adjacent higher ability group. Note that these small groups were excluded from the calculation of conditional covariance estimates used in Zhang and Stout (1999a). This process repeats until the number of examinees in the merged group exceeds the lower bound. It is recommended that this lower bound be at least 10. After that, the sample covariance of two items, denoted as $\widehat{\text{cov}}(X_{i_1}, X_{i_2} \mid \text{Group } k)$, is computed using response data from Group k only. Then, a sample expected conditional covariance can be constructed as

$$\widehat{E\text{cov}}_{i_1 i_2}(T) = \sum_k \frac{J_k}{J} \widehat{\text{cov}}(X_{i_1}, X_{i_2} \mid \text{Group } k),$$

where J is the total number of examinees, J_k is the number of examinees in group k for $k = 1, 2, \ldots, K$, and K is the number of final valid examinees' groups.

Similarly, the *remaining score* or *rest score* can also be used as a conditioning variable to estimate conditional covariances. For a fixed item pair (i_1, i_2), the remaining score is $S_{i_1 i_2} = \sum_{i=1, i \neq i_1, i_2}^{n} X_i$. The sample expected conditional covariance, using the remaining score as a conditioning variable, can be constructed as

$$\widehat{E\text{cov}}_{i_1 i_2}(S) = \sum_k \frac{J_{i_1 i_2 k}}{J} \widehat{\text{cov}}(X_{i_1}, X_{i_2} \mid \text{Group } k \text{ based on } S_{i_1 i_2}),$$

where $J_{i_1 i_2 k}$ is the number of examinees of Group k based on the remaining score $S_{i_1 i_2}$, and $\widehat{\text{cov}}(X_{i_1}, X_{i_2} \mid \text{Group } k \text{ based on } S_{i_1 i_2})$ is the sample covariance of Group k.

The final estimator of expected conditional covariance used in this paper is in the same form as that used by Zhang and Stout (1999a) for dichotomous items. That is,

$$\widehat{E\text{cov}}^*_{i_1 i_2} = \tfrac{1}{2} \left[\widehat{E\text{cov}}_{i_1 i_2}(S) + \widehat{E\text{cov}}_{i_1 i_2}(T) \right]. \tag{20}$$

The purpose of using the average of the two estimators is to reduce the bias (see Zhang & Stout, 1999a). Generally, one may optimally choose a combination weight w between 0 and 1 such that $w \widehat{E\text{cov}}_{i_1 i_2}(S) + (1 - w) \widehat{E\text{cov}}_{i_1 i_2}(T)$ has minimal bias. Yang and Zhang (2001) investigated this issue. Their results show that the estimator (20) with $w = 0.5$ is near-optimal.

The estimators of expected conditional correlation may calculated as

$$\widehat{E\text{Cor}}_{i_1 i_2}(S) = \frac{\widehat{E\text{cov}}_{i_1 i_2}(S)}{\sqrt{\widehat{E\text{cov}}_{i_1 i_1}(S)\, \widehat{E\text{cov}}_{i_2 i_2}(S)}} \quad \text{and} \quad \widehat{E\text{Cor}}_{i_1 i_2}(T) = \frac{\widehat{E\text{cov}}_{i_1 i_2}(T)}{\sqrt{\widehat{E\text{cov}}_{i_1 i_1}(T)\, \widehat{E\text{cov}}_{i_2 i_2}(T)}}.$$

Then, the final estimator of expected conditional correlation is

$$\widehat{E\text{Cor}}^*_{i_1 i_2} = \frac{1}{2} \left[\widehat{E\text{Cor}}_{i_1 i_2}(S) + \widehat{E\text{Cor}}_{i_1 i_2}(T) \right]. \tag{21}$$

The major advantage of the use of an observed score as a conditional variable is that a composite score does not need to be estimated. However, when total raw scores are not available because of missing data by design, this approach may not be applicable and a composite score should be used instead as a conditioning variable.

4. Justification of DETECT for Polytomous Items

DETECT, short for *dimensionality evaluation to enumerate contributing traits*, is a statistical procedure that is used to identify the number of dominant latent dimensions and to estimate the degree of multidimensionality (see Zhang & Stout, 1999b). DETECT can correctly assign items to dimensionally homogeneous clusters when approximate simple structure exists.

4.1. The Theoretical DETECT Index

The definition of the theoretical DETECT index is derived from the properties of expected conditional covariances presented in Section 2. If a test is split into K nonempty and disjoint sets of items, say, $A_1, A_2, \ldots, A_K$, then $\mathcal{P} = \{A_1, A_2, \ldots, A_K\}$ is called a *K-subset partition* of the test. The DETECT index (Zhang & Stout, 1999b) is defined as

$$D(\mathcal{P}) = \frac{2}{n(n-1)} \sum_{1 \le i_1 < i_2 \le n} \delta_{i_1 i_2}(\mathcal{P}) E[\mathrm{cov}(X_{i_1}, X_{i_2} \mid \Theta_\alpha)], \tag{22}$$

where $\mathcal{P}$ is any partition of a test, Θ_α is a given composite, especially the test composite, and

$$\delta_{i_1 i_2}(\mathcal{P}) = \begin{cases} 1, & \text{if items } X_{i_1} \text{ and } X_{i_2} \text{ are in the same subset,} \\ -1, & \text{otherwise.} \end{cases} \tag{23}$$

Although originally defined for dichotomous items, the theoretical DETECT index remains exactly the same in form for polytomous items (including dichotomous items as special cases). There are $n(n-1)/2$ terms in the summation of (22). Hence, the index is, in fact, an algebraic average of all expected conditional covariances of item pairs. Note that the expected conditional covariances may be replaced by the expected conditional correlation coefficients so that different types of items have relatively even contributions to the index. Here, the expected conditional correlation is defined as

$$E\mathrm{Cor}(X_{i_1}, X_{i_2} \mid \alpha) = \frac{E[\mathrm{cov}(X_{i_1}, X_{i_2} \mid \Theta_\alpha)]}{\sqrt{E[\mathrm{Var}(X_{i_1} | \Theta_\alpha)] E[\mathrm{Var}(X_{i_2} | \Theta_\alpha)]}}. \tag{24}$$

Given a partition $\mathcal{P}$, the expected conditional covariance, $E[\mathrm{cov}(X_{i_1}, X_{i_2} \mid \Theta_\alpha)]$, is either added or subtracted in (22) depending on whether items X_{i_1} and X_{i_2} come from the same subset in the partition $\mathcal{P}$ or not. Let

$$M^* = \frac{2}{n(n-1)} \sum_{1 \le i_1 < i_2 \le n} |E[\mathrm{cov}(X_{i_1}, X_{i_2} \mid \Theta_\alpha)]|.$$

M^* is an upper bound of the theoretical DETECT index by its definition. Suppose a multidimensional test has an approximate simple structure. Let $\mathcal{P}^*$ be the dimensionality-based cluster partition (i.e., the partition that matches the existing approximate simple structure). Then, $D(\mathcal{P}^*) = M^*$, that is, $D(\cdot)$ achieves its maximum value M^* at $\mathcal{P}^*$. Moreover, $\mathcal{P}^*$ is the unique partition that maximizes the index $D(\cdot)$ because any other partitions of the test will reduce the magnitude of $D(\cdot)$. The main idea of the DETECT procedure is to search for the partition that maximizes an estimate of the theoretical DETECT index. This partition is regarded as the dimensionality-based cluster partition $\mathcal{P}^*$. It can be expected that a well-estimated DETECT index will also be maximized at $\mathcal{P}^*$ if there is sufficient examinee data to guarantee its statistical accuracy. In Section 6 simulation studies will be conducted to check the performance of an estimated DETECT index defined in Section 4.3.

As discussed after Theorem 2 in Section 2, if two items are dimensionally homogeneous, the magnitude of the expected conditional covariance of these two items indicates the degree of dimensional homogeneity of these two items; that is, the larger the magnitude, the greater the degree of dimensional homogeneity. Otherwise, if two items are dimensionally heterogeneous, the magnitude of the expected conditional covariance indicates the degree of dimensional heterogeneity of these two items; that is, the larger the absolute value of the expected conditional covariance, the greater the degree of dimensional heterogeneity. Therefore, the magnitude of the maximum DETECT value indicates the degree of multidimensionality the test displays (i.e., the size of the departure from being perfectly fitted by a unidimensional model). As Zhang and Stout (1999b) pointed out, this index will often be useful from the perspective of statistical robustness, for example, when assessing the appropriateness of using BILOG (Mislevy & Bock, 1982) or PARSCALE, which presumes unidimensionality.

4.2. The Performance of the Theoretical DETECT for Polytomous Items

For dichotomously scored items, the theoretical DETECT index has been proven to be maximized at the correct cluster partition of a test with approximate simple structure, where each cluster in this partition corresponds to a distinct dominant dimension under certain reasonable conditions (Zhang & Stout, 1999b). Since the same conditional covariance results have been established for polytomous items as for dichotomous items in Section 2, all results of DETECT for dichotomous items also hold for polytomous items under the two assumptions given in Section 2, which are also assumed in the remainder of this section. The proofs of theorems for polytomous items are similar to those for dichotomous items. Thus, this paper presents below theorems without detailed proofs.

When a test is unidimensional, any composite is the latent trait variable Θ itself. Because items are pairwise locally uncorrelated, $\text{cov}(X_{i_1}, X_{i_2} \mid \Theta) = 0$ for all $1 \leq i_1 < i_2 \leq n$. Thus, $D(\mathcal{P}) = 0$ for any partition $\mathcal{P}$. Conversely, when $D(\mathcal{P}) = 0$ for any partition $\mathcal{P}$, it is not difficult to prove that $E[\text{cov}(X_{i_1}, X_{i_2} \mid \Theta_\alpha)] = 0$ for all $1 \leq i_1 < i_2 \leq n$. According to Theorem 1, $\text{cov}(X_{i_1}, X_{i_2} | \Theta_\alpha = \theta) = 0$ for all θ. That is, these items are pairwise locally uncorrelated with respect to Θ_α and hence are unidimensional. Therefore, a test is unidimensional if, and only if, $D(\mathcal{P}) = 0$ for any partition $\mathcal{P}$.

If a test is two dimensional, according to Corollary 1, the theoretical DETECT $D(\mathcal{P})$ will be maximized at a two-cluster partition of the test. Further, each of the two clusters will be composed of the items on the same side of the composite direction, and every item will be uniquely in one of two clusters except those items whose discrimination parameter vectors have exactly the same direction as the composite Θ_α. The conditional covariances involving those items are zero, and hence, those items can be in either cluster since they have no contribution to the DETECT value. Those items should be few in real operational situations. Hence we have the following results.

Theorem 3. *When the number of dimensions of a test is one or two, the theoretical DETECT index can always count the dimensions correctly and identify a two-cluster solution for a two-dimensional case.*

When the number of dimensions exceeds two, this paper mainly considers tests with approximate simple structure.

Theorem 4. *If a d-dimensional test has approximate simple structure, then the theoretical DETECT will be maximized uniquely at the K-cluster partition $\mathcal{P}^*$ satisfying* (17) *with $K \leq d$.*

Note that d is the number of dimensions according to the mathematical definition of dimensionality. Zhang and Stout (1999b) argued that the $K < d$ situation is likely to happen, and in such cases the K is a more appropriate number than d to describe the dimensional structure of a test; that is, in some heuristic sense, K is the number of dominant dimensions.

In practice, simple structure is widely assumed in the analysis of response data. Mathematically, a test is called a *simple structure test* if there exists a d-dimensional latent coordinate system such that all the items lie along the coordinate axes and there is at least one item along each axis. Tests are often designed to display such simple structures approximately when their frameworks require that each item simply measure one of several separate subscales. It should be noted that the simple structure coordinate system is allowed to be oblique in the sense that any two subscales are correlated, that is, $\text{cov}(\Theta_i, \Theta_j) > 0$ for all $1 \le i < j \le d$. Clearly, there are d distinct clusters corresponding to the d subscales in such a simple structure test. Let A_j be the jth cluster (one-scale subtest) and suppose there are n_j items in A_j. Thus, $\sum_{j=1}^{d} n_j = n$. Clearly, $\mathcal{P}^* = \{A_1, A_2, \ldots, A_d\}$ is the correct d-cluster partition of the test. One needs to know when the theoretical DETECT is maximized at the partition $\mathcal{P}^*$. Note that a simple structure test is a special case of approximate simple structure except when there exists two items such that $\beta_{i_1 i_2} = \pi/2$, that is, (17) does not hold.

The *regularity condition* defined by Zhang and Stout (1999b) is needed in the following theorem. The intuitive meaning of the regularity condition is that there are no overly dominant correlation coefficients between any two subscales relative to the correlation coefficients between all subscale pairs. Note that the regularity condition always holds for any two-dimensional simple structure test.

Theorem 5. *For a simple structure test, the theoretical DETECT is maximized uniquely at the dimensionally correct d-cluster partition $\mathcal{P}^*$ if, and only if, the regularity condition holds. Otherwise, it will be maximized at a K-cluster partition $\mathcal{P}_0 = \{B_1, B_2, \ldots, B_K\}$, where $K < d$ and each B_k is the union of some A_j's for $k = 1, 2, \ldots, K$.*

Once again, K can be interpreted as the number of dominant dimensions because if the regularity condition does not hold, some A_j's must be very close (i.e., the respective abilities they load on are highly correlated); these A_j's form a large cluster B_k that measures one dominant dimension. In such a case, the partition $\mathcal{P}_0 = \{B_1, B_2, \ldots, B_K\}$ satisfies (17).

To better understand Theorem 5, consider a three-dimensional simple structure test with approximately the same number of items in each subscale (so that the test composite has equal weights on all three subscales). If one subscale is only moderately correlated with the other two subscales, then the regularity condition holds when the other two subscales are not too highly correlated. For example, if $\rho_{12} = \rho_{13} = 0.70$, then the regularity condition holds when $\rho_{23} < 0.88$, but will not hold when $\rho_{23} > 0.88$ (see Table 2 in Zhang & Stout, 1999b). In the former case, the theoretical *DETECT* is maximized uniquely at $\mathcal{P}^* = \{A_1, A_2, A_3\}$, while in the latter case, it is maximized uniquely at $\mathcal{P}_0 = \{A_1, B_2\}$ with $B_2 = A_2 \cup A_3$, according to Theorem 5. Readers may consider the above example as a verbal and math test with 20 verbal, 20 algebra, and 20 geometry items. If the correlation between algebra and geometry turns out to be too high, then the theoretical DETECT will be maximized at the two-cluster partition with 20 verbal and 40 math items as its two clusters. In some sense, it is a reasonable solution. Of course, the DETECT may be further applied to the 40-item math subtest, and theoretically it will find the partition with algebra and geometry clusters.

Zhang and Stout (1999b) proposed two theoretical indexes: One is called the *approximate simple structure index* and the other is the *ratio index*, denoted as ASSI($\mathcal{P}$) and R($\mathcal{P}$), respectively. Both can be developed into statistical indexes for judging whether a response data set displays approximate simple structure or not. For any partition $\mathcal{P}$ of a test, define

$$\text{ASSI}(\mathcal{P}) = \frac{2}{n(n-1)} \sum_{1 \le i_1 < i_2 \le n} \delta_{i_1 i_2}(\mathcal{P}) \, \text{Sgn}(E[\text{cov}(X_{i_1}, X_{i_2} \mid \Theta_T)]) \tag{25}$$

and

$$R(\mathcal{P}) = \frac{\sum_{1 \le i_1 < i_2 \le n} \delta_{i_1 i_2}(\mathcal{P}) E[\text{cov}(X_{i_1}, X_{i_2} \mid \Theta_T)]}{\sum_{1 \le i_1 < i_2 \le n} \left| E[\text{cov}(X_{i_1}, X_{i_2} \mid \Theta_T)] \right|}, \tag{26}$$

where Θ_T is a test composite and $\delta_{i_1i_2}(\mathcal{P})$ is defined by (23). These two indexes range from -1 to $+1$. Hence, they may be regarded as standardized versions of the DETECT index. Like the DETECT index, they have the same form for both dichotomous and polytomous items.

Theorem 6.

(1) *A test has approximate simple structure if, and only if,* $\max_{\mathcal{P}} \mathrm{ASSI}(\mathcal{P}) = 1$.
(2) *A test has approximate simple structure if, and only if,* $\max_{\mathcal{P}} R(\mathcal{P}) = 1$.

Therefore, the magnitudes of both indexes are the indicators of the presence of approximate simple structure. One can also define these two indexes based on the conditional correlation coefficients. Note that the ASSI($\mathcal{P}$) based on conditional correlation coefficients is exactly the same as that based on conditional covariances according to (24).

4.3. Estimation of DETECT and the DETECT Procedure

Section 3 discussed how to estimate the expected conditional covariance, and two types of estimators were presented there. After obtaining an estimator $\widehat{E\mathrm{cov}}_{i_1i_2}$, either $\widehat{E\mathrm{cov}}_{i_1i_2}(\alpha)$ in (18) or $\widehat{E\mathrm{cov}}^{*}_{i_1i_2}$ in (20), it is easy to construct an estimator of the theoretical DETECT by substituting the expected conditional covariances with their corresponding estimators, that is,

$$\widehat{D}(\mathcal{P}) = \frac{2}{n(n-1)} \sum_{1\le i_1<i_2\le n} \delta_{i_1i_2}(\mathcal{P})\, \widehat{E\mathrm{cov}}_{i_1i_2},$$

where $\delta_{i_1i_2}(\mathcal{P})$ is defined by (23). Similarly, one can construct estimators for ASSI($\mathcal{P}$) and $R(\mathcal{P})$, that is,

$$\widehat{\mathrm{ASSI}}(\mathcal{P}) = \frac{2}{n(n-1)} \sum_{1\le i_1<i_2\le n} \delta_{i_1i_2}(\mathcal{P})\, \mathrm{Sgn}(\widehat{E\mathrm{cov}}_{i_1i_2})$$

and

$$\widehat{R}(\mathcal{P}) = \frac{\sum\limits_{1\le i_1<i_2\le n} \delta_{i_1i_2}(\mathcal{P})\, \widehat{E\mathrm{cov}}_{i_1i_2}}{\sum\limits_{1\le i_1<i_2\le n} \left|\widehat{E\mathrm{cov}}_{i_1i_2}\right|}.$$

Similarly, one can use an estimator of conditional correlation $\widehat{E\mathrm{Cor}}_{i_1i_2}$ given by (19) or (21) to construct estimators of DETECT indexes.

A computer program, called PolyDETECT, was developed based on the procedure proposed in this paper. PolyDETECT reports two separate parts of results: one is based on conditional covariances and the other on conditional correlations. Depending on the nature of the test, one of the two parts, or both, may be used to determine the dimensional structure of response data. Currently, the estimates given by (20) and (21) are used in PolyDETECT by default. Note that the original DETECT software only used an old version of conditional covariances (20) for dichotomous items.

The operating rule of PolyDETECT, which remains almost the same as the original DETECT program, is to search for a partition that maximizes an estimated DETECT index and judge that partition, called the optimal partition, to be the dimensional-based cluster partition. The search engine for the optimal partition is a genetic algorithm (Zhang & Stout, 1999b), which remains exactly the same for response data with polytomous items. When forming the optimal partition on the basis of estimated conditional covariances/correlations, it is possible that statistical noise dominates the searching process, especially in unidimensional cases where the optimal partition is formed solely due to statistical noise. To prevent the error of failure to detect the unidimensionality of a test, an additional sample of examinees is needed to perform cross-validation. In practice, response data are usually randomly divided into two parts with roughly the same size (e.g., a 40×1000 data set with 40 items and 1000 examinees is randomly divided into two 40×500 data sets) whenever possible. Both half-data sets are used to calculate the estimated DETECT indexes, $\widehat{D}_1(\cdot)$ and $\widehat{D}_2(\cdot)$, and to search for their respective optimal partitions, $\mathcal{P}_1^*$ and $\mathcal{P}_2^*$, independently. If $\mathcal{P}_1^*$ and $\mathcal{P}_2^*$ are approximately the same, then these optimal partitions are considered to be formed due to the intrinsic dimensional structure of the response data. Otherwise, the two optimal partitions are regarded to be formed by capitalization upon chance. In PolyDETECT, this is accomplished by comparing the reference DETECT value with the maximum DETECT value. Here, the maximum and reference DETECT values refer to the DETECT values using the first half data at $\mathcal{P}_1^*$ and $\mathcal{P}_2^*$. That is, $\widehat{D}_{\text{max}} = \widehat{D}_1(\mathcal{P}_1^*)$ and $\widehat{D}_{\text{ref}} = \widehat{D}_1(\mathcal{P}_2^*)$. Theoretically, $\widehat{D}_{\text{ref}}$ is less than or equal to $\widehat{D}_{\text{max}}$. If $\mathcal{P}_1^*$ and $\mathcal{P}_2^*$ are exactly the same, then $\widehat{D}_{\text{ref}}$ equals $\widehat{D}_{\text{max}}$. If $\widehat{D}_{\text{ref}}$ is significantly smaller than $\widehat{D}_{\text{max}}$ (i.e., $\mathcal{P}_1^*$ and $\mathcal{P}_2^*$ are quite different from each other), one may suspect that the optimal partitions $\mathcal{P}_1^*$ and $\mathcal{P}_2^*$ are not formed due to the test's intrinsic multidimensional structure, but by statistical noise. If $\widehat{D}_{\text{ref}}$ is near zero, or even negative, then the test under investigation can be inferred to be essentially unidimensional.

When determining the dimensional structure, the other two indexes are also used. Only when $\widehat{\text{ASSI}}_1(\mathcal{P}_2^*)$ or $\widehat{\text{R}}_1(\mathcal{P}_2^*)$ is larger than a critical value will PolyDETECT declare that the data set is multidimensional. The critical values currently selected as default are 0.3 and 0.4 for these two indexes, respectively. When both $\widehat{\text{ASSI}}_1(\mathcal{P}_2^*)$ and $\widehat{\text{R}}_1(\mathcal{P}_2^*)$ are smaller than some critical values (default values are 0.25 and 0.36, respectively), the program may declare the data set essentially unidimensional. Otherwise, PolyDETECT will ask its user to use other information to determine the dimensional structure of response data.

When PolyDETECT declares a data set multidimensional, the number of *sizable* clusters in the partition $\mathcal{P}_1^*$ is judged as the number of dominant dimensions present in the test as Zhang and Stout (1999b) proposed. The sizable clusters are those that contain at least a certain number of items or a certain proportion of the test. The current default for the lower bound in PolyDETECT is $n/13$ truncated to be between 3 and 9, where n is the number of items in the test.

As pointed out before, many tests, such as the Graduate Record Examinations® (GRE®) General Test, TOEFL®, NAEP main assessments, and so forth, are composed of several sections or subsets of items measuring different subscales. It is important to know whether or not the statistical dimensional structure is in concert with the substantive test structure. One way to check that is to calculate the DETECT value at the content-based partition and then compare this value with the maximum DETECT value. If they are relatively close to each other, then one would say the content-based partition is near-optimal. The PolyDETECT program lets its user provide such a partition for a confirmatory analysis.

Various types of errors can happen in the applications of PolyDETECT. One type of possible error is that PolyDETECT incorrectly enumerates the number of dimension(s). Another is that PolyDETECT assigns some items to wrong clusters when it correctly identifies the number of dimensions. In the next section, simulation studies will be conducted to check the error rates in various cases.

5. Simulation Studies

In this section, simulation studies were conducted to check the performance of PolyDETECT based on the estimators given by (20) and (21). In the simulation studies, response data sets were generated to be either unidimensional or multidimensional with simple structure. Then PolyDETECT was applied to these data sets in an effort to recover their dimensional structure.

The estimated item parameters from the analysis of the 2002 NAEP Grades 4 and 8 reading assessments were used as true item parameters to generate simulated data sets. According to the contemporary definition of reading literacy, the NAEP reading items were developed in accordance with three contexts for reading and four aspects of reading (see National Assessment Governing Board, 1992). The three contexts for reading are *reading for literary experience*, *reading to gain information*, and *reading to perform a task*. NAEP assesses all three contexts for reading in Grade 8, but only the first two contexts in Grade 4. Each cognitive item belongs to one, and only one, context for reading, and each context-based subtest (i.e., all items related to the same context) is considered to be unidimensional.

The number of items in the 2002 NAEP Grade 4 reading assessment was 82; each context for reading had 41 items. While for Grade 8, there were 111 items with 30, 48, and 33 items measuring the three contexts for reading, respectively. There are two types of items in NAEP assessments: multiple-choice and constructed-response items. The multiple-choice items are scored dichotomously, but some of the constructed-response items may be scored polytomously if they require somewhat more elaborate responses. The test composition is listed in Table 1.

There is a "bad" item in the second subscale for each grade. Both are multiple-choice items with large difficulty and lower-asymptote parameters (e.g., Item 71 in Grade 4 with $b = 3.025$, and $c = 0.312$). These two items were excluded from the simulation studies. Thus, the total number of items for Grades 4 and 8 in the simulation is 81 and 110, respectively. The test length considered in this simulation study is 20, 40, 60, or 81 for Grade 4 cases, and 45, 60, 90, or 110 for Grade 8 cases. A test with less than 81 items in Grade 4 cases, or less than 110 items in Grade 8 cases, consists of an equal number of items in each subscale. For example, a 40-item test for Grade 4 consists of 20 items from the first subscale and another 20 items from the second subscale.

The total number of examinees in each PolyDETECT run with cross validation is 1000, 2000, 4000, 6000, 8000, or 10,000, with half as a target data set and the other half as reference. Examinees' (true) ability scores were generated independently from a multivariate normal distribution with means of 0, variances of 1, and a common correlation coefficient of 0.0, 0.3, 0.6, 0.8, 0.9, or 1.0. When the correlation is one, subscales are the same and the corresponding

TABLE 1.
NAEP reading test composition by subscale and item type.

	Literary	Gain information	Perform a task	Total
Grade 4				
MC	18	19	NA	37
CR-D	16	11	NA	27
CR-P	7	11	NA	18
Total	41	41	NA	82
Grade 8				
MC	12	19	11	42
CR-D	8	14	16	38
CR-P	10	15	6	31
Total	30	48	33	111

Note: MC stands for multiple-choice items, and CR-D and CR-P stand for constructed-response items scored dichotomously and polytomously, respectively.

cases are unidimensional. Thus, simulated Grade 4 response data are either two dimensional or unidimensional, and Grade 8 data are either three dimensional or unidimensional.

Given the three factors, the number of items, the number of examinees, and the correlation coefficient between subscales, there are 144 combinations for each grade. For each combination, the simulation and analysis process (generating a response data set and applying PolyDETECT to identify its dimensional structure) was replicated 100 times.

The results of simulation studies are summarized in Tables 2 and 3 for Grades 4 and 8, respectively. Because of length limitations, this paper only reports the number of times out of 100 replications that PolyDETECT correctly identified the number of dimension(s) and the number of times that PolyDETECT further correctly assigned items into dimensionally based clusters for multidimensional cases (i.e., the optimal partition of items found by PolyDETECT was exactly the true dimensionally based one). Tables 2 and 3 present these two counts obtained by PolyDETECT based on both (20) and (21) in order to compare the performance of PolyDETECT based on conditional correlations with that based on conditional covariances. Hence, there are four counts for a multidimensional case and two counts for a unidimensional case in Tables 2 and 3. When these four, or two for unidimensional cases, numbers happen to be the same in a cell, only one number is presented. The first number in each cell is the count out of 100 replications for which PolyDETECT correctly declared the number of dimensions based on (20), and the second is the corresponding count based on (21). For multidimensional cases, the third number is the count for which PolyDETECT not only correctly declared the number of dimensions, but also identified the true dimensionally based partition based on (20), and the fourth number is the corresponding count based on (21).

Tables 2 and 3 show that PolyDETECT found the true dimensionally based partition in every replication when the correlation was low or moderate and the number of examinees was large. PolyDETECT completely recovered the true dimensional structure of response data in all 100 replications in 98 (Grade 4) and 85 (Grade 8) cases out of the total of the 144 cases for each grade in the simulation study. PolyDETECT successfully identified the unidimensional cases in every replication in every case except for the cases of 20 items and 4000 or more examinees. Generally, when the number of items is small, the statistical error of the estimate of conditional covariance/correlation is relatively large since the classification of examinees into homogeneous ability groups based on the total test score may not be very reliable. In this situation, if the number of examinees is large, systematic errors (e.g., bias) driven by different item characteristics may arise and dominate the magnitude of the estimate in unidimensional cases. Hence, PolyDETECT based on the default estimates (20) and (21) can be used only when the test length is not too short. When the correlation is 0.9 and the total number of examinees is 1000, PolyDETECT does not seem to perform well: Most times PolyDETECT either could not determine the dimensional structure or it incorrectly declared the test to be essentially unidimensional. For example, when the number of items is 60, the number of examinees is 1000, and the correlation coefficient is 0.9, the frequency with which PolyDETECT based on conditional covariances or correlations declared response data sets to be two dimensional is only 14 or 12 out of 100 replications (see Table 2). In that case, it may be reasonable to claim the test to be essentially unidimensional since the correlation is so high. Nevertheless, the overall rates from Tables 2 and 3 that PolyDETECT correctly declared the number of dimension(s) are 96% and 95% for grades 4 and 8, respectively.

The results in Tables 2 and 3 also indicate that the performance of PolyDETECT based on conditional correlations (21) is almost the same as that based on conditional covariances (20) in the situations considered in this simulation study. Although not all DETECT statistics are reported here, it should be noted that in unidimensional cases reference values are significantly smaller than maximum values, and reference values are near zero or even negative except for the cases of 20 items and 4000 or more examinees. The maximum DETECT value has a negative association

Table 2.
Frequency of (completely) correct DETECT results out of 100 replications using conditional covariance/correlation in unidimensional or two-dimensional cases (Grade 4).

		Number of examinees					
n	ρ	1000	2000	4000	6000	8000	10,000
20	0.0	100	100	100	100	100	100
	0.3	100	100	100	100	100	100
	0.6	100, 100 (100, 99)	100	100	100	100	100
	0.8	100, 99 (72, 68)	100, 100 (94, 92)	100	100	100	100
	0.9	22, 18 (6, 9)	79, 81 (35, 41)	100, 100 (84, 85)	100, 100 (88, 91)	100, 100 (95, 94)	100, 100 (99,97)
	1.0	100	100	95, 90	81, 77	58, 46	41, 26
40	0.0	100	100	100	100	100	100
	0.3	100	100	100	100	100	100
	0.6	100, 100 (98, 97)	100	100	100	100	100
	0.8	98, 97 (54, 53)	100, 100 (84, 84)	100, 100 (96, 96)	100, 100 (99, 98)	100, 100 (100, 99)	100
	0.9	11, 11 (3, 3)	91, 94 (25, 27)	100, 100 (63, 58)	100, 100 (71, 78)	100, 100 (81, 77)	100, 100 (86,85)
	1.0	100	100	100	100	100	100
60	0.0	100	100	100	100	100	100
	0.3	100	100	100	100	100	100
	0.6	100, 100 (95, 95)	100	100	100	100	100
	0.8	100, 100 (46, 44)	100, 100 (88, 87)	100, 100 (99, 99)	100, 100 (99, 100)	100	100
	0.9	14, 12 (4, 3)	94, 94 (26, 23)	100, 100 (66, 65)	100, 100 (85, 87)	100, 100 (87, 89)	100, 100 (93, 94)
	1.0	100	100	100	100	100	100
81	0.0	100	100	100	100	100	100
	0.3	100	100	100	100	100	100
	0.6	100, 100 (94, 94)	100, 100 (99, 99)	100	100	100	100
	0.8	100, 100 (56, 54)	100, 100 (90, 92)	100	100	100	100
	0.9	14, 9 (2, 1)	99, 99 (22, 25)	100, 100 (80, 79)	100, 100 (92, 92)	100, 100 (97, 97)	100
	1.0	100	100	100	100	100	100

Note. There are four numbers in each cell in the table. If all four numbers in a cell are the same, then only one number is presented in that cell. n is the number of items, ρ is the population correlation coefficient between subscales, and $\rho = 1.0$ represents a unidimensional case.

TABLE 3.
Frequency of (completely) correct DETECT results out of 100 replications using conditional covariance/correlation in unidimensional or three-dimensional cases (Grade 8).

		Number of examinees					
n	ρ	1000	2000	4000	6000	8000	10,000
45	0.0	100, 100 (99, 99)	100	100	100	100	100
	0.3	100, 100 (95, 95)	100	100	100	100	100
	0.6	100, 100 (76, 76)	100, 100 (97, 95)	100, 100 (99, 99)	100	100	100
	0.8	96, 93 (15, 18)	100, 100 (51, 55)	100, 100 (90, 88)	100, 100 (93, 94)	100, 100 (97, 97)	100, 100 (96, 96)
	0.9	0	30, 12 (2, 1)	99, 100 (19, 17)	100, 100 (35, 31)	100, 100 (39, 44)	100, 100 (44, 53)
	1.0	100	100	100	100	100	100
60	0.0	100	100	100	100	100	100
	0.3	100, 100 (98, 97)	100	100	100	100	100
	0.6	100, 100 (81, 79)	100, 100 (99, 98)	100	100	100	100
	0.8	99, 97 (9, 6)	100, 100 (53, 52)	100, 100 (93, 93)	100, 100 (94, 95)	100, 100 (98, 99)	100, 100 (99, 99)
	0.9	0	37, 25 (3, 1)	100, 100 (28, 24)	100, 100 (41, 37)	100, 100 (50, 53)	100, 100 (62, 71)
	1.0	100	100	100	100	100	100
90	0.0	100	100	100	100	100	100
	0.3	100, 100 (95, 96)	100	100	100	100	100
	0.6	100, 100 (76, 79)	100, 100 (99, 99)	100	100	100	100
	0.8	100, 100 (11, 6)	100, 100 (69, 64)	100, 100 (97, 95)	100, 100 (98, 98)	100	100
	0.9	0	73, 47 (4, 4)	100, 100 (37, 41)	100, 100 (74, 73)	100, 100 (80, 80)	100, 100 (88, 89)
	1.0	100	100	100	100	100	100
110	0.0	100	100	100	100	100	100
	0.3	100, 100 (96, 95)	100	100	100	100	100
	0.6	100, 100 (80, 76)	100, 100 (99, 99)	100	100	100	100
	0.8	100, 100 (10, 10)	100, 100 (56, 55)	100, 100 (98, 98)	100, 100 (98, 99)	100, 100 (99, 99)	100
	0.9	0	50, 21 (2, 4)	100, 100 (38, 39)	100, 100 (81, 78)	100, 100 (89, 87)	100, 100 (97, 96)
	1.0	100	100	100	100	100	100

Note. There are four numbers in each cell in the table. If all four numbers in a cell are the same, then only one number is presented in that cell. n is the number of items, ρ is the population correlation coefficient between subscales, and $\rho = 1.0$ represents a unidimensional case.

with the correlation coefficients between subscales. The smaller the correlation coefficients, the larger the maximum DETECT value, implying greater departure from unidimensionality. This suggests that the magnitude of the maximum DETECT value is informative in indicating the degree of multidimensionality in the test displays, and the correlation of the underlying abilities (subscales) is one of the important factors in determining the degree of multidimensionality.

6. Applying PolyDETECT to NAEP Reading Data

In this section, the PolyDETECT program is used to analyze the dimensional structure of the 2002 NAEP reading Grades 4 and 8 operational data.

Because of the test time limitation, no one student takes all the items in the NAEP assessments. A matrix-sampling design of test items, called the *focused balanced incomplete block* (BIB) *spiraling* design, has been implemented in the main NAEP assessments (Beaton, Johnson, & Ferris, 1987). Reading passages and accompanying items are divided into blocks. Each sampled student is given a test booklet typically containing two 25-minute blocks of items or one 50-minute block. In the 2002 NAEP reading assessment of Grade 8, for example, there were a total of 10 different blocks (one 50-minute and nine 25-minute blocks) and 37 different booklets (36 booklets from the BIB design of nine 25-minute blocks plus the 50-minute block booklet). Missing data are treated according to the NAEP conventions in this study (see Allen, Donoghue, & Schoeps, 2001).

The reporting scales of NAEP reading assessments are based on the contexts for reading in the reading operational analysis conducted at the Educational Testing Service (ETS). Separate IRT-based subscales have been developed for each of the contexts for reading, and the methodology of multiple imputations (plausible values) is used to estimate key population features. A composite that is a weighted average of the plausible values of all the subscales is then created as a measure of overall proficiency. The weight for each reading subscale is the target proportion of items measuring that context. For example, the weights of grade 8 for the three contexts for reading are 0.4, 0.4, and 0.2, respectively. For details, see Allen et al. (2001). As discussed in Section 3, this composite can be used as the conditioning variable when calculating conditional covariances/correlations.

Although the classification of items into context-based subscales has substantive meaning, statistical justification is needed for this analysis approach. One natural question is whether this simple multiple-subscale structure based on the contexts for reading reflects well the structure of the NAEP reading data. That is, is such a classification optimal in the sense that items in the same cluster are relatively dimensionally homogenous while items from different clusters are not? Or is there another classification of items that better matches the structure of the data than the context-based one does? For example, does the reading-aspect-based or an item-type-based partition better match the structure of the data than the context-based partition?

This study uses the reporting samples of the 2002 NAEP reading assessment of Grades 4 and 8 with sample sizes 139,383 and 114,681, respectively. Each sample is randomly split into two parts of roughly the same size to run the PolyDETECT program with cross validation. Recall that the PolyDETECT program reports both results based on conditional covariances and on conditional correlations. The DETECT results based on conditional correlations turned out to be the same as those based on conditional covariances for both grades. Therefore, only the results based on conditional covariances are presented here.

The DETECT results show that the fourth-grade data set is two dimensional. The optimal two-cluster partition of items provided by PolyDETECT is

$$\{\{1\text{–}41, 71\}, \{42\text{–}70, 72\text{–}82\}\},$$

which is consistent with the substantive two-cluster partition determined by the NAEP reading design,

$$\{\{1\text{–}41\}, \{42\text{–}82\}\},$$

except for one multiple-choice item, Item 71. The discrimination, difficulty, and lower-asymptote parameter estimates of this item from the 2002 NAEP operational analysis are 0.597, 3.025, and 0.312, respectively, in the proficiency scale of mean zero and variance one. Clearly, this is a very difficult item with a high lower-asymptote (guessing) parameter and is not a good item from the psychometric point of view (low item information).

PolyDETECT concludes that the eighth-grade response data set is three dimensional; its optimal three-cluster partition is

$$\{\{1\text{–}30, 40\}, \{31\text{–}39, 41\text{–}65\}, \{66\text{–}111\}\}.$$

This partition agrees with the substantive partition based on three contexts for reading,

$$\{\{1\text{–}30\}, \{31\text{–}78\}, \{79\text{–}111\}\},$$

except for item 40 and the 50-minute block (Items 66–78). Item 40 is a multiple-choice item with large difficulty and lower-asymptote parameters ($b = 2.151$ and $c = 0.319$). It should be noted that the booklet with the 50-minute block was separate/independent from the BIB design in the NAEP assessment. For any item pair with one item from the 50-minute block and the other from any other block, the conditional covariance is not estimable since no one student completed such a pair of items. In cases where no students completed a pair of items, the PolyDETECT program automatically assigns a zero value to the conditional covariance, which indicates that there is no information about the degree of dimensional homogeneity of these two items. Therefore, the 50-minute block can be moved among clusters without changing or affecting the value of DETECT. That is, the 50-minute block as a whole can be put into any context-based cluster without changing or affecting the value of DETECT, though all items in that block are related to the context for reading to gain information. Consistently, PolyDETECT always keeps all of the 50-minute block items in the same cluster, indicating that these items are relatively dimensionally homogeneous given the composite.

Table 4 presents the three index values at the three partitions from the PolyDETECT program. The three indexes reported here are the DETECT index, the approximate simple structure index (ASSI), and the ratio index (R). The three partitions presented here are the optimal partition obtained from the target data set (labeled as Maxima), the optimal partition obtained from the reference data set (labeled as Reference), and the context-based partition (labeled as Context-based). For each grade, these three partitions are the same except for one or two items and for the additional 50-minute block for Grade 8. For Grade 4, the DETECT values at the three partitions are approximately the same, which indicates the outlier item has little contribution to the DETECT value. For Grade 8, the values of the three indexes at the context-based partition are the same as those at the reference partition (the last two rows of Table 4 for Grade 8). Their only difference is the 50-minute block, which does not affect these index values at all. In addition, these values are only slightly smaller than their corresponding values at the optimal partition. For both grades, the values of ASSI are relatively small, which indicates that the 2002 NAEP reading data for Grades 4 and 8 are weakly multidimensional, which is most likely due to the high correlations between subscales. Moreover, this fact may also indicate that the simple structure assumption is too strong for these data sets. Overall, the index values are very close across three partitions for each grade, indicating that the context-based partition of items is valid and optimal under the assumption of approximate simple structure.

TABLE 4.
DETECT results for the 2002 NAEP reading data for Grades 4 and 8.

	Grade 4			Grade 8		
Partition	DETECT	ASSI	R	DETECT	ASSI	R
Maxima	0.0173	0.2864	0.6258	0.0133	0.2039	0.6357
Reference	0.0173	0.2821	0.6235	0.0132	0.2033	0.6337
Context-based	0.0173	0.2809	0.6239	0.0132	0.2033	0.6337

7. Discussion

In this paper the theory of conditional covariances originally developed for dichotomous items is extended to polytomous items. The theory provides a theoretical foundation for procedures based on conditional covariances/correlations, such as DETECT and DIMTEST, so that the performance of these procedures is theoretically justified when applied to response data with polytomously scored items. Two types of estimators of conditional covariances are constructed and discussed. With these estimators of conditional covariances, the DETECT procedure can be applied not only to response data sets with polytomous items, but also to complex sampling data sets with missing values, either by design or at random. PolyDETECT can be applied to verify whether the content-based classification of items into clusters (subtests) is statistically consistent with the dimensional structure of the data through exploratory and confirmatory analysis. Simulation studies show that PolyDETECT performs well with a large sample and with balanced numbers of items in various dimensions. Further studies are still needed to investigate the performance of PolyDETECT when the sample size is small, or when the numbers of items in various dimensions are extremely unbalanced.

In this paper, PolyDETECT was applied to analyze the dimensional structure of the 2002 NAEP reading samples of Grades 4 and 8. Zwick (1987) assessed the dimensionality of the 1983–1984 NAEP reading data. Her conclusion is that “it is not unreasonable to treat that data as unidimensional.” Although there have been great changes in the NAEP reading assessment since then, PolyDETECT indicates that the 2002 NAEP reading data sets are weakly multidimensional. At the same time, the DETECT results in this study also show that the context-based partition of items into clusters is optimal if items need to be classified according to their multidimensional structure. Yu and Nandakumar (2001) applied DETECT to analyze the 1992 NAEP eighth-grade reading data and conclude that the data set has “at most moderate degree of multidimensionality.” Their dimensionality analysis was carried out at test booklet and selected item-subset levels since the early version of DETECT was not yet capable of handling whole BIB-designed data. The composites corresponding to the conditional variables used in their analysis may be quite different from each other in different booklet-level runs because of NAEP BIB design. Consequently, the degree of multidimensionality might be overestimated.

Although the new version of PolyDETECT works well in the simulation studies, it is still an open question as to how to construct a consistent unbiased estimator for $E[\mathrm{cov}(X_i, X_j \mid \Theta_T)]$ so as to obtain a good estimator of the theoretical DETECT index, especially when the test length is short. Another research topic is to establish a large sample distribution theory for DETECT. That is, one needs to understand the statistical behavior of DETECT so that statistical hypothesis testing can be carried out for testing whether a response data set is d-dimensional or not. All these issues are still under investigation.

References

Allen, N., Carlson, J.E., & Zelenak, C. (1999). *The NAEP 1996 technical report* (NCES 1999-452). Washington, DC: Office of Educational Research and Improvement, US Department of Education.

Allen, N., Donoghue, J.R., & Schoeps, T.L. (2001). *The NAEP 1998 technical report* (NCES 2001-509). Washington, DC: Office of Educational Research and Improvement, US Department of Education.

Anderson, T.W. (1984). *An introduction to multivariate statistical analysis* (2nd ed.). New York: Wiley.

Beaton, A.E., Johnson, E.G., & Ferris J.J. (1987). The assignment of exercises to students. In A.E. Beaton (Ed.), *Implementing the new design: The NAEP 1983–84 technical report* (pp. 97–118). Princeton, NJ: Educational Testing Service.

Douglas, J., Kim, H.R., & Stout, W.F. (1994). Exploring and explaining the lack of local independence through conditional covariance functions. Paper presented at the 1994 (April) annual meeting of the American Educational Research Association, New Orleans, LA.

Habing, B., & Roussos, L.A. (2003). On the need for negative local item dependence. *Psychometrika, 68*, 435–451.

Holland, P.W., & Rosenbaum, P.R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics, 14*, 1523–1543.

Junker, B. (1993). Conditional association, essential independence and monotone unidimensional item response models. *The Annals of Statistics, 21*, 1359–1378.

Kim, H.R. (1994). *New techniques for the dimensionality assessment of standardized test data*. Unpublished doctoral dissertation, Department of Statistics, University of Illinois at Urbana-Champaign.

McDonald, R.P. (1994). Testing for approximate dimensionality. In D. Laveault, B. Zumbo, M. Gessaroli, & M. Boss (Eds.), *Modern theories of measurement: Problems and issues* (pp. 63–85). Ottawa: University of Ottawa Press.

Mislevy, R., & Bock, R.D. (1982). *BILOG: Item analysis and test scoring with binary logistic models [Computer software]*. Mooresville, IN: Scientific Software.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159–176.

Muraki, E., & Bock, R.D. (1997). *PARSCALE: IRT item analysis and test scoring for rating scale data [Computer software]*. Chicago: Scientific Software.

Nandakumar, R., & Stout, W.F. (1993). Refinement of Stout's procedure for assessing latent trait essential unidimensionality. *Journal of Educational Statistics, 18*, 41–68.

National Assessment Governing Board. (1992). *Reading framework for the National Assessment of Educational Progress: 1992–2002*. Washington, DC: National Assessment Governing Board.

National Assessment Governing Board. (2002). *Mathematics framework for the 2003 National Assessment of Educational Progress*. Washington, DC: National Assessment Governing Board.

Oltman, P.K., Stricker, L.J., & Barrows, T.S. (1990). Analyzing test structure by multidimensional scaling. *Journal of Applied Psychology, 75*, 21–27.

Reckase, M.D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9*, 401–412.

Reckase, M.D., & McKinley, R.L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15*, 361–373.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. Psychometrika Monograph, No. 17. Greensboro, NC: Psychometric Society.

Samejima, F. (1972). *A general model for free-response data*. Psychometrika Monograph No. 18. Greensboro, NC: Psychometric Society.

Stout, W.F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika, 52*, 589–617.

Stout, W.F., Habing, B., Douglas, J., Kim, H.R., Roussos, L.A., & Zhang, J. (1996). Conditional covariance based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 20*, 331–354.

Van Abswoude, A.A.H., Van der Ark, L.A., & Sijtsma, K. (2004). A comparative study on test dimensionality assessment procedures under nonparametric IRT models. *Applied Psychological Measurement, 28*, 3–24.

Wang, M. (1986). *Fitting a unidimensional model on the multidimensional item response data* (ONR Technical Report 87-1). Iowa City, IA: University of Iowa.

Yang, X., & Zhang, J. (2001). Construction and evaluation of bias-corrected estimators of DETECT dimensionality index. Paper presented at the 2001 (April) annual meeting of the American Educational Research Association, Seattle, WA.

Yu, F., & Nandakumar, R. (2001). Poly-DETECT for quantifying the degree of multidimensionality of item response data. *Journal of Educational Measurement, 38*, 99–120.

Zhang, J. (1996). *Some fundamental issues in item response theory with applications*. Unpublished doctoral dissertation, Department of Statistics, University of Illinois at Urbana-Champaign.

Zhang, J., & Stout, W.F. (1999a). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika, 64*, 129–152.

Zhang, J., & Stout, W.F. (1999b). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64*, 213–249.

Zwick, R. (1987). Assessing the dimensionality of NAEP reading data. *Journal of Educational Measurement, 24*, 293–308.

Manuscript received 25 OCT 2004
Final version received 10 JAN 2006
Published Online Date: 30 JAN 2007