Research Report

Conditional Covariance Theory and DETECT for Polytomous Items

Jinming Zhang

Research & Development

December 2004
RR-04-50

Conditional Covariance Theory and DETECT for Polytomous Items

Jinming Zhang

ETS, Princeton, NJ

December 2004

Abstract

This paper extends the theory of conditional covariances to polytomous items. It has been mathematically proven that under some mild conditions, commonly assumed in the analysis of response data, the conditional covariance of two items, dichotomously or polytomously scored, is positive if the two items are dimensionally homogeneous and negative otherwise. 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 NAEP data. 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 by 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. The results also indicate that the degree of multidimensionality of the NAEP reading data is weak.

Key words: Item response theory, IRT, multidimensional item response theory, MIRT, dimensionality, multidimensionality, approximate simple structure, cluster analysis.

Acknowledgments

This report is an extension and revision of the paper presented at the annual meeting of American Educational Research Association, Chicago, Illinois, April 1997. The research was supported by Educational Testing Service and the National Assessment of Educational Progress (Grant R902F980001), U.S. Department of Education. The opinions expressed herein are solely those of the author and do not necessarily represent those of Educational Testing Service. The author would like to thank Ting Lu, Paul Holland, Shelby Haberman, and Feng Yu for their comments and suggestions.

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. 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. 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 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.

Several statistical methods are available for dimensionality analysis: factor analysis, cluster analysis, and multidimensional scaling. 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*. "Basically, the factor model is motivated by the following argument. Suppose variables can be grouped by their correlations. That is, all variables within a particular group are highly correlated among themselves but have relatively small correlations with variables in a different group. It is conceivable that each group of variables represents a single underlying construct, or factor, that is responsible for the observed correlations" (see Johnson & Wichern, 1992, pp. 396-397). 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 et al., 1996; Zhang & Stout, 1999a; Habing & Roussos, 2003). The expected conditional covariance is

$$E[\mathrm{Cov}(X_{i_1}, X_{i_2}|Y)] = \sum_k P(Y=k)\mathrm{Cov}(X_{i_1}, X_{i_2}|Y=k),$$

where Y is an appropriately chosen observed score (e.g., the total raw score). To better understand the performance of these procedures, researchers studied the behavior of $\mathrm{Cov}(X_{i_1}, X_{i_2}|Y=k)$ or $E[\mathrm{Cov}(X_{i_1}, X_{i_2}|Y)]$. Rosenbaum (1984), Holland and Rosenbaum (1986), Junker (1993), and Douglas et al. (1994) investigated the properties of $\mathrm{Cov}(X_{i_1}, X_{i_2}|Y=k)$ when a test is unidimensional. However, it is too difficult to study its properties directly for multidimensional cases. Instead of studying $\mathrm{Cov}(X_{i_1}, X_{i_2}|Y=k)$, Zhang and Stout (1999a) investigated the structure and the properties of $\mathrm{Cov}(X_i, X_j|\Theta_Y=\theta)$ for dichotomously scored items, where Θ_Y is an appropriately chosen composite best measured by score Y. A composite is a standardized linear combination of the latent trait variables. When the test length is long enough, $E[\mathrm{Cov}(X_{i_1}, X_{i_2}|Y)]$ should be close to $E[\mathrm{Cov}(X_{i_1}, X_{i_2}|\Theta_Y)]$. That is, the conditional covariance given an observed score can be regarded as an approximation to conditional covariance given an appropriate value of the composite that is best measured by the observed score. Thus, researchers may study the properties of $E[\mathrm{Cov}(X_{i_1}, X_{i_2}|Y)]$ by investigating the properties of $E[\mathrm{Cov}(X_{i_1}, X_{i_2}|\Theta_Y)]$. Zhang and Stout (1999a) proved that if a test has an approximate simple structure, the conditional covariance given Θ_Y will be positive when two items come from the same cluster and negative when two items come from different clusters. 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 provides a solid theoretical foundation for recently developed dimensionality assessment procedures such as DETECT (Kim, 1994; Zhang, 1996) and DIMTEST (Stout, 1987; Nandakumar & Stout, 1993), which

are conditional covariance based procedures. 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.

The purposes of this paper are to study the 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. Then, the theory of conditional covariances developed in Section 2 is used to theoretically justify the DETECT procedure for a test with polytomous items in Section 4. In Section 5, DETECT is used to analyze the 2002 NAEP reading data. Section 6 presents some simulation results, and 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. Multidimensional item response theory (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 $\boldsymbol{\Theta} = (\Theta_1, \Theta_2, \ldots, \Theta_d)'$, where $\boldsymbol{\Theta}$ 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 $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_d)'$. That is,

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

$P_{ik}(\boldsymbol{\theta})$ 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,

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

When the item is dichotomously scored, $F_i(\boldsymbol{\theta}) = P_{i1}(\boldsymbol{\theta}) = P(X_i = 1|\boldsymbol{\theta})$. Thus, an IRF is an extension of the item response function of a dichotomous item. It is assumed that an IRF is (monotone) increasing, that is, the expected score of an item increases monotonically when at least one of the abilities increases. It is also assumed that *local independence* holds, that is, $X_1, X_2, \ldots, X_n$ are independent given $\boldsymbol{\Theta}$. Some researchers (McDonald, 1994; Stout et al., 1996) suggest using a weak version of local independence, that is,

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

for all $\boldsymbol{\theta}$ and $1 \leq i_1 < i_2 \leq n$. In this paper, this weak version of local independence is adopted and called the *pairwise local independence*.

A test is said to be *d-dimensional* if d is the minimal number of abilities required to produce a pairwise local independent, monotone latent model. When $d = 1$, the test is called *unidimensional*.

A *composite*, Θ_α, of the latent vector $\boldsymbol{\Theta}$ is defined to be a standardized linear combination of $\boldsymbol{\Theta}$, that is,

$$\Theta_\alpha = \boldsymbol{\alpha}'\boldsymbol{\Theta} = \sum_{j=1}^{d} \alpha_j \Theta_j$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_d)'$ is any fixed constant vector and 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 mathematically 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 (see Zhang & Stout, 1999a). It should be noted that these two assumptions are sufficient but not necessary conditions, that is, conditional covariances may still have these properties even if these two assumptions do not hold completely.

2.1 Two Assumptions

The first assumption is that the latent trait vector $\boldsymbol{\Theta}$ has a multivariate normal distribution, $\boldsymbol{\Theta} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = (\rho_{ij})$ 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 $\boldsymbol{\Sigma}$ 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

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

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*. 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 the other dimensions. It 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 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 generalized compensatory model also includes the multidimensional compensatory versions of the generalized partial credit model and the graded response model. Hence, almost all commonly used compensatory MIRT models are special cases of the generalized compensatory model.

The ICRF of a generalized partial credit model for a unidimensional case (Muraki,

1992) is

$$P_{ik}(\theta) = P(X_i = k \mid \theta) = \frac{\exp\{\sum_{v=0}^{k} 1.7a_i(\theta - b_i + d_{iv})\}}{\sum_{j=0}^{m_i} \exp\{\sum_{v=0}^{j} 1.7a_i(\theta - b_i + d_{iv})\}} \tag{4}$$

for $k = 0, 1, \ldots, m_i$, where a_i, b_i, and d_{ik} are unknown item parameters, $d_{i0} = 0$ and $\sum_{k=1}^{m_i} d_{ik} = 0$. There are $m_i + 1$ independent item parameters in an item with $m_i + 1$ categories. This model can be rewritten as

$$P_{ik}(\theta) = \frac{\exp\{1.7((k+1)a_i\theta - b_{ik})\}}{\sum_{j=0}^{m_i} \exp\{1.7((j+1)a_i\theta - b_{ij})\}} \tag{5}$$

where $b_{ik} = a_i((k+1)b_i - \sum_{v=0}^{k} d_{iv})$ for $k = 0, 1, \ldots, m_i$, and $b_{im_i} = (m_i + 1)b_{i0}$. The ICRF of a multidimensional compensatory generalized partial credit model is defined as

$$P_{ik}(\boldsymbol{\theta}) = P(X_i = k \mid \boldsymbol{\theta}) = \frac{\exp\{(k+1)\mathbf{a}_i'\boldsymbol{\theta} - b_{ik}\}}{\sum_{j=0}^{m} \exp\{(j+1)\mathbf{a}_i'\boldsymbol{\theta} - b_{ij}\}} \quad \text{for} \quad k = 0, 1, \ldots, m_i, \tag{6}$$

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

$$F_i(\boldsymbol{\theta}) \equiv E[X_i|\boldsymbol{\theta}] = H_{i1}(\mathbf{a}_i'\boldsymbol{\theta}),$$

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

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

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 (6) 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) can be written as

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

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

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

The corresponding IRF can be obtained

$$F_i(\boldsymbol{\theta}) \equiv E[X_i|\boldsymbol{\theta}] = \sum_{k=1}^{m_i} P_{ik}^*(\boldsymbol{\theta}) = H_{i2}(\mathbf{a}_i^{'}\boldsymbol{\theta}),$$

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 is also a special case of a generalized compensatory model.

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 $\Theta_\alpha = \boldsymbol{\alpha}^{'}\boldsymbol{\Theta}$, define

$$\lambda_{i_1 i_2} = \mathrm{Cov}(\mathbf{a}_{i_1}^{'}\boldsymbol{\Theta}, \mathbf{a}_{i_2}^{'}\boldsymbol{\Theta} \mid \Theta_\alpha), \tag{7}$$

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),

$$\lambda_{i_1 i_2} = \mathbf{a}_{i_1}^{'}\boldsymbol{\Sigma}\,\mathbf{a}_{i_2} - \frac{(\mathbf{a}_{i_1}^{'}\boldsymbol{\Sigma}\,\boldsymbol{\alpha})(\mathbf{a}_{i_2}^{'}\boldsymbol{\Sigma}\,\boldsymbol{\alpha})}{\boldsymbol{\alpha}^{'}\boldsymbol{\Sigma}\,\boldsymbol{\alpha}}. \tag{8}$$

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

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

$$\mathrm{Sgn}[\mathrm{Cov}(X_{i_1}, X_{i_2} \mid \Theta_\alpha)] = \mathrm{Sgn}(\lambda_{i_1 i_2}), \tag{9}$$

where $\mathrm{Sgn}(x)$ *is the sign function that gives the sign (+, −, or 0) of* x*. That is, for all* θ*,* $\lambda_{i_1 i_2}$ *and* $\mathrm{Cov}(X_{i_1}, X_{i_2} \mid \Theta_\alpha = \theta)$ *always have the same sign,*

$$\mathrm{Cov}(X_{i_1}, X_{i_2} \mid \Theta_\alpha = \theta) \begin{cases} > 0, & \text{if } \lambda_{i_1 i_2} > 0; \\ = 0, & \text{if } \lambda_{i_1 i_2} = 0; \\ < 0, & \text{if } \lambda_{i_1 i_2} < 0. \end{cases}$$

Moreover, $\mathrm{Cov}(X_{i_1}, X_{i_2} \mid \Theta_\alpha = \theta)$ *is a strictly increasing function of* $\lambda_{i_1 i_2}$ *when* $d > 2$*, and* $\lambda_{i_1 i_1}$ *and* $\lambda_{i_2 i_2}$ *are fixed. These results also hold for the expected conditional covariance* $E[\mathrm{Cov}(X_{i_1}, X_{i_2} \mid \Theta_\alpha)]$.

Proof. By the conditional covariance formula, and then by local independence,

$$\begin{aligned} \mathrm{Cov}(X_{i_1}, X_{i_2} \mid \Theta_\alpha) &= \mathrm{Cov}(E(X_{i_1}|\boldsymbol{\Theta}), E(X_{i_2}|\boldsymbol{\Theta}) \mid \Theta_\alpha) + E(\mathrm{Cov}(X_{i_1}, X_{i_2}|\boldsymbol{\Theta}) \mid \Theta_\alpha) \\ &= \mathrm{Cov}[F_{i_1}(\boldsymbol{\Theta}), F_{i_2}(\boldsymbol{\Theta}) \mid \Theta_\alpha], \end{aligned}$$

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

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

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), (9) 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 $\lambda_{i_1 i_2}$ when $d > 2$, and $\lambda_{i_1 i_1}$ and $\lambda_{i_2 i_2}$ are fixed.

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

$$E[\mathrm{Cov}(X_{i_1}, X_{i_2}|\Theta_\alpha)] = \int_{-\infty}^{\infty} \mathrm{Cov}(X_{i_1}, X_{i_2}|\Theta_\alpha = \theta) f_{\boldsymbol{\alpha}}(\theta) d\theta \tag{10}$$

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 (9) holds for any given θ value. Therefore,

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

and $E[\mathrm{Cov}(X_{i_1}, X_{i_2} \mid \Theta_\alpha)]$ is a strictly increasing function of $\lambda_{i_1 i_2}$ when $d > 2$, and $\lambda_{i_1 i_1}$ and $\lambda_{i_2 i_2}$ 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 following corollary is obtained using the same proof of Corollary 2 in Zhang and Stout (1999a).

Corollary 1. *Suppose* $d = 2$ *and* $\Theta_\alpha = \boldsymbol{\alpha}'\boldsymbol{\Theta}$ *is any given composite. Then for any given* θ*,*

$$\mathrm{Cov}(X_{i_1}, X_{i_2}|\Theta_\alpha = \theta) \begin{cases} > 0, & \text{if the discrimination parameter vectors of items} \\ & i_1 \text{ and } i_2 \text{ are on the same side of vector } \boldsymbol{\alpha}; \\ = 0, & \text{if at least one of the discrimination parameter} \\ & \text{vectors is in the same direction as vector } \boldsymbol{\alpha}; \\ < 0, & \text{if the discrimination parameter vectors of items} \\ & i_1 \text{ and } i_2 \text{ are on different sides of vector } \boldsymbol{\alpha}; \end{cases} \tag{11}$$

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

$$< \mathbf{a}_i, \mathbf{a}_j >= \mathbf{a}_i'\boldsymbol{\Sigma}\,\mathbf{a}_j$$

for any $\mathbf{a}_i, \mathbf{a}_j \in V$, where $\boldsymbol{\Sigma}$ is the correlation matrix of the latent trait vector $\boldsymbol{\Theta}$. Then V is a d-dimensional Euclidean space. The length of vector $\mathbf{a}$ is defined as $\|\mathbf{a}\| \equiv \sqrt{< \mathbf{a}, \mathbf{a} >}$, and the angle β between $\mathbf{a}_i$ and $\mathbf{a}_j$ is defined as

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

Let

$$V_{\boldsymbol{\alpha}}^{\perp} = \{\mathbf{a} \in V : \ < \mathbf{a}, \boldsymbol{\alpha} >= 0\}. \tag{12}$$

$V_{\boldsymbol{\alpha}}^{\perp}$ that is orthogonal to $\boldsymbol{\alpha}$ is a $(d-1)$-dimensional subspace of d-dimensional Euclidean space V. 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* $\Theta_\alpha = \boldsymbol{\alpha}'\boldsymbol{\Theta}$*,*

$$\lambda_{i_1 i_2} = \|\mathbf{a}_{i_1}^{\perp}\| \, \|\mathbf{a}_{i_2}^{\perp}\| \cos \beta_{i_1 i_2} \tag{13}$$

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

$$\mathrm{Cov}(X_{i_1}, X_{i_2} | \Theta_\alpha = \theta) = 0, \tag{14}$$

and

$$E[\mathrm{Cov}(X_{i_1}, X_{i_2} | \Theta_\alpha)] = 0. \tag{15}$$

Otherwise,

$$\mathrm{Cov}(X_{i_1}, X_{i_2} | \Theta_\alpha = \theta) \begin{cases} > 0, & \text{if } \beta_{i_1 i_2} < \pi/2; \\ = 0, & \text{if } \beta_{i_1 i_2} = \pi/2; \\ < 0, & \text{if } \beta_{i_1 i_2} > \pi/2; \end{cases} \tag{16}$$

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* $E[Cov(X_{i_1}, X_{i_2} \mid \Theta_\alpha)]$ *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

$$\mathbf{a}_{i_1} = c_1 \boldsymbol{\alpha} + \mathbf{a}_{i_1}^{\perp} \tag{17}$$

$$\mathbf{a}_{i_2} = c_2 \boldsymbol{\alpha} + \mathbf{a}_{i_2}^{\perp} \tag{18}$$

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 (8), (17), and (18), one can obtain (13). By Theorem 1 and (13), one obtains (14) and (16). Since $\lambda_{i_1 i_1} = \|\mathbf{a}_{i_1}^{\perp}\|^2$ and $\lambda_{i_2 i_2} = \|\mathbf{a}_{i_2}^{\perp}\|^2$, 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 (10), the results for expected conditional covariance are obtained. □

Note that $\|\mathbf{a}_i^{\perp}\| = \|\mathbf{a}_i\| \sin \gamma_i$ where γ_i is the angle formed by $\mathbf{a}_i$ and $\boldsymbol{\alpha}$. Fixing $\|\mathbf{a}_i^{\perp}\|$ does not mean fixing $\mathbf{a}_i$ or $\|\mathbf{a}_i\|$. For example, $\mathbf{a}_i$ still can rotate along the cone with axis $\boldsymbol{\alpha}$ and angle γ_i. Further, $\mathbf{a}_i$ may come out of the cone (i.e., γ_i may change) as long as the length of $\mathbf{a}_i$ also changes accordingly so that $\|\mathbf{a}_i\| \sin \gamma_i$ remains constant. Theorem 2 shows that the closer the item pair directions are (i.e., the greater the degree of dimensional homogeneity of that item pair is), 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. Note that dimensionally homogeneous or heterogeneous means that the items are close or are not close to each other in $V_{\boldsymbol{\alpha}}^{\perp}$, the subspace of the latent space, regardless of their original directions in the latent trait space. That is, one judges whether two items are dimensionally homogeneous or not by examining whether their directions are close to each other or not relative to the composite direction.

Since the (expected) conditional covariances have nice properties related to the dimensional structure of a test, the similarity based on the (expected) conditional covariances is a good candidate to be used in cluster analysis or multidimensional scaling to analyze the dimensionality of a test. Suppose that a test consists of relatively dimensionally distinct clusters with each cluster consisting of relatively dimensionally homogeneous items. The major objective of dimensionality analysis is to discover this true cluster partition of test items. If an appropriate composite, which should be in the center of all the item discrimination vectors or the item directions, can be chosen as a conditioning variable, then by Theorem 2, any two items in the same cluster are positively conditionally correlated, whereas any two items from different clusters are negatively conditionally correlated. Therefore, the test items can be grouped into several distinct clusters such that any two items will be in the same cluster if, and only if, the expected conditional covariance of these two items is positive. Further, the number of clusters may be judged to be the number

of dominant dimensions present in the test, and the cluster in which an item is located corresponds to the dominant dimension that the item is measuring. Hence, under the assumptions of Section 2.1 one can find the true dimensional structure of the test by using the expected conditional covariances. The next section discusses how to appropriately choose a manifest variable as a conditioning variable to form sample conditional covariances. Starting with Section 4, this paper will use the theory developed here to justify the DETECT procedure when applied to response data with polytomous items.

3. Sample Conditional Covariance

The previous section presented some important properties of (expected) conditional covariances that can be utilized to analyze dimensional structure of response data. However, this cannot be done until an accurate estimator of conditional covariance is found.

3.1 A Composite Scale Score as a Conditioning Variable

In this approach, an appropriate composite is estimated for each examinee. The idea is that a unidimensional calibration program, such as PARSCALE, 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: unidimensional approximation approach and 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). Intuitively, the direction of the reference composite will be in the center of all the item directions (i.e., a centroid of item directions). The simple structure approach regards the whole response data as multidimensional with simple structure. Since an educational test usually has intentionally concentrated measurement subscales, the simple structure assumption is widely used in practice. 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 maximum raw scores of subtests or even weights.

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. The number of groups is recommended to be between 11 and 50. For example, in the analysis of the 2002 NAEP reading data in Section 6, the number of groups is 20 and the 5th, 10th, ..., 95th percentiles (with 5% increments) are used as cut-points for grouping examinees. 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 all sample item pair covariances in the usual way, that is,

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

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 $E[\mathrm{Cov}(X_{i_1}, X_{i_2} | \Theta_\alpha)]$ by

$$\widehat{\mathrm{Cov}_{i_1 i_2}}(\boldsymbol{\alpha}) = \sum_k \frac{J_k}{J} \widehat{\mathrm{Cov}}(X_{i_1}, X_{i_2} \mid \text{ Group } k), \tag{19}$$

where the $\boldsymbol{\alpha}$ 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

$$\widehat{\mathrm{Corr}_{i_1 i_2}}(\boldsymbol{\alpha}) = \frac{\widehat{\mathrm{Cov}_{i_1 i_2}}(\boldsymbol{\alpha})}{\sqrt{\widehat{\mathrm{Cov}_{i_1 i_1}}(\boldsymbol{\alpha}) \widehat{\mathrm{Cov}_{i_2 i_2}}(\boldsymbol{\alpha})}}. \tag{20}$$

In NAEP, although no scale scores are reported for individual students, posterior means of students for subscales are estimated during NAEP operational analysis. Posterior means are estimators of students' subscale scores. Thus a composite of posterior means can be used as a conditioning variable in the calculation of conditional covariances/correlations. If posterior means are not available, plausible values (see Mislevy, Johnson, & Muraki, 1992) may also be used in place of scale scores. Under such circumstances, special attention

should be paid to the accuracy of conditional covariance estimates due to the sparseness of response data.

3.2 An Observed Raw Score as a Conditioning Variable

As discussed in Section 1, researchers would focus on conditional covariances given an observed raw score. Obstacles to the theoretical study of conditional covariances given an observed score have led to the investigation of conditional covariances given a composite. In practice, however, 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 composite 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 Reckase and McKinley (1991) and Zhang and Stout (1999a). This composite and the reference composite discussed in Section 3.1 are generally referred to as the test composite in this paper.

For the sake of simplicity, take the total score, $T = \sum_{i=1}^{n} X_i$, as an example to illustrate how to form sample conditional covariance given an observed score. First, examinees are stratified into several groups according to their total scores. Then, a sample covariance is obtained for each group. If the perfect (highest) total score is N ($N \geq n$), 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. 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. This process repeats until the number of examinees in the merged group exceeds the lower bound. It is recommended that this lower bound be set between 10 and 50. After that, the sample covariance of two items, denoted as $\widehat{\mathrm{Cov}}(X_{i_1}, X_{i_2}|\text{ Group } k)$, is computed using response data from group k only. Finally, a sample conditional covariance can be constructed as

$$\widehat{\mathrm{Cov}_{i_1 i_2}}(T) = \sum_k \frac{J_k}{J} \widehat{\mathrm{Cov}}(X_{i_1}, X_{i_2}|\text{ Group } k),$$

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.

Note that an item with (m+1) categories, scored as $0, 1, \ldots, m$, may have up to m points contributing to the total raw score. For example, a three-category item may have a two-point contribution while a dichotomous item may only have a one-point contribution. If equal maximum contribution of all items to the total score is preferred in some special cases, then the weighted score, $Y_w = \sum_{i=1}^{n} X_i/m_i$, can be used as a conditioning variable, and the (initial) k-th group of examinees are those with $k - 1 < Y_w \leq k$, $k = 0, 1, 2, \ldots, n$.

Kim (1994) uses the *remaining score* as a conditioning variable to estimate conditional covariances. For a fixed item pair (i_1, i_2), the remaining score is $S_{i_1 i_2} = \sum_{i=1, i \neq i_1, i_2}^{n} X_i$. Examinees are partitioned into groups according to their remaining scores. Then, the sample conditional covariance, using the remaining score as a conditioning variable, can be constructed as

$$\widehat{\mathrm{Cov}}_{i_1 i_2}(S) = \sum_{k} \frac{J_{i_1 i_2 k}}{J} \widehat{\mathrm{Cov}}(X_{i_1}, X_{i_2} | \text{ Group } k \text{ based on } S_{i_1 i_2}),$$

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 $\widehat{\mathrm{Cov}}(X_{i_1}, X_{i_2} |$ Group k based on $S_{i_1 i_2})$ is the sample covariance of group k.

When a test is unidimensional, Rosenbaum (1984), Holland and Rosenbaum (1986) show that $\mathrm{Cov}(X_{i_1}, X_{i_2} | S_{i_1 i_2}) > 0$ and Junker (1993) proves that $\mathrm{Cov}(X_{i_1}, X_{i_2} | T) \leq 0$ for the Rasch model, whereas $\mathrm{Cov}(X_{i_1}, X_{i_2} | \Theta) = 0$. As a result, the sample conditional covariance using the total score as conditioning variable could have a negative bias, whereas the sample conditional covariance using the remaining score as conditioning variable could have a positive bias. Zhang and Stout (1999a) suggest combining these two estimators to reduce the bias. Thus, the final sample conditional covariance is

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

Yang and Zhang (2001) investigate how to choose the combination weights to maximally reduce the bias. Their results show that this estimator is near optimal. Tables 1 and 2 of Zhang and Stout (1999a) show that this final estimator is much closer to zero than

$\widehat{\text{Cov}}_{i_1 i_2}(S)$ and $\widehat{\text{Cov}}_{i_1 i_2}(T)$ for unidimensional dichotomous items. This is also true for unidimensional polytomous items. For multidimensional items, the sign behavior of this estimator is consistent with that of the expected conditional covariance if the number of examinees is large enough, according to simulation studies. For details, see Yang and Zhang (2001). The estimators of expected conditional item pair correlation may calculated as

$$\widehat{\text{Corr}}_{i_1 i_2}(S) = \frac{\widehat{\text{Cov}}_{i_1 i_2}(S)}{\sqrt{\widehat{\text{Cov}}_{i_1 i_1}(S)\widehat{\text{Cov}}_{i_2 i_2}(S)}} \quad \text{and} \quad \widehat{\text{Corr}}_{i_1 i_2}(T) = \frac{\widehat{\text{Cov}}_{i_1 i_2}(T)}{\sqrt{\widehat{\text{Cov}}_{i_1 i_1}(T)\widehat{\text{Cov}}_{i_2 i_2}(T)}}.$$

Then, the final estimator of expected conditional correlation is

$$\widehat{\text{Corr}}^{*}_{i_1 i_2} = \frac{1}{2}\left[\widehat{\text{Corr}}_{i_1 i_2}(S) + \widehat{\text{Corr}}_{i_1 i_2}(T)\right]. \tag{22}$$

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 conditional variable. For a test with a complex matrix design, such as NAEP, the latter approach may be the only feasible way to get an overall 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, to estimate the degree of multidimensionality, and/or to 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

$$D(\mathcal{P}) = \frac{2}{n(n-1)} \sum_{1 \le i_1 < i_2 \le n} \delta_{i_1 i_2}(\mathcal{P}) E[\mathrm{Cov}(X_{i_1}, X_{i_2} | \Theta_\alpha)] \tag{23}$$

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

$$\delta_{i_1 i_2}(\mathcal{P}) = \begin{cases} 1, & \text{if items } X_{i_1} \text{ and } X_{i_2} \text{ are in the same subset,} \\ -1, & \text{otherwise.} \end{cases} \tag{24}$$

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 (23). 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

$$\mathrm{Corr}(X_{i_1}, X_{i_2} | \boldsymbol{\alpha}) = \frac{E[\mathrm{Cov}(X_{i_1}, X_{i_2} | \Theta_\alpha)]}{\sqrt{E[\mathrm{Cov}(X_{i_1}, X_{i_1} | \Theta_\alpha)] E[\mathrm{Cov}(X_{i_2}, X_{i_2} | \Theta_\alpha)]}}. \tag{25}$$

Given a partition $\mathcal{P}$, the expected conditional covariance, $E[\mathrm{Cov}(X_{i_1}, X_{i_2} | \Theta_\alpha)]$, is either added or subtracted in (23) depending on whether items X_{i_1} and X_{i_2} come from the same subset in the partition $\mathcal{P}$ or not. Let

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

From (23), M^* is an upper bound of the theoretical DETECT index. If a multidimensional test exhibits approximate simple structure, then the expected conditional covariance will be positive when the two items come from the same dimensionally homogeneous cluster; it will be negative otherwise, according to the theory of conditional covariance presented in Section 2. 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. Section 6 will show the results of simulation studies to check the performance of an estimated DETECT index defined later.

As discussed in Section 2 just after Theorem 2, if two items come from the same dimensionally homogeneous cluster, 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. If two items come from dimensionally different clusters, the magnitude of the expected conditional covariance of these two items indicates the degree of dimensional heterogeneity of these two items; that is, the larger the magnitude, 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 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 mathematically 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. Note that the theoretical DETECT is

a nonparametric index that does not require any particular parametric forms for the item response functions and the distribution of the latent trait vector. Here these two assumptions need to demonstrate mathematically the performance of this theoretical index. This paper conjectures that the results of DETECT below as well as the results of conditional covariances presented in Section 2 will still hold for other reasonable models such as multidimensional noncompensatory models. The proofs of theorems for polytomous items are similar to those for dichotomous items. Thus, this paper presents theorems below without detailed proofs.

For a unidimensional test, any composite is the latent trait variable Θ itself. Because of the local independence, $\mathrm{Cov}(X_{i_1}, X_{i_2}|\Theta) \equiv 0$ for all $1 \leq i_1 < i_2 \leq n$. Thus, $E[\mathrm{Cov}(X_{i_1}, X_{i_2}|\Theta)] \equiv 0$ and $D(\mathcal{P}) \equiv 0$ for a unidimensional test. Conversely, when $D(\mathcal{P}) \equiv 0$ for any partition $\mathcal{P}$, it is not difficult to prove that $E[\mathrm{Cov}(X_{i_1}, X_{i_2}|\Theta_\alpha)] = 0$ for all $1 \leq i_1 < i_2 \leq n$. According to Theorem 1, $\mathrm{Sgn}[\mathrm{Cov}(X_{i_1}, X_{i_2}|\Theta_\alpha = \theta)] = \mathrm{Sgn}[E[\mathrm{Cov}(X_{i_1}, X_{i_2}|\Theta_\alpha)]] = 0$ for all θ. That is, $\mathrm{Cov}(X_{i_1}, X_{i_2}|\Theta_\alpha = \theta) = 0$ for all θ; or the weak local independence holds. Therefore, if $D(\mathcal{P}) \equiv 0$ for any partition $\mathcal{P}$, then the test is unidimensional.

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 unique 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.

Theorem 3. *When the number of dimensions of a test is one or two, the theoretical DETECT index can always count the test 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. Informally, a test is said to have approximate simple structure if the test consists of relatively dimensionally distinct clusters with each cluster consisting of relatively dimensionally homogeneous items. Formally, a test has approximate simple structure if there exists a partition $\mathcal{P}^* = \{A_1, A_2, \ldots, A_K\}$ such that the conditional covariances given a test composite are positive for every item pair from the same A_k, and negative for every item pair from the different A_ks. Geometrically, a test has approximate simple structure if, and only if, there exists a partition $\mathcal{P}^* = \{A_1, A_2, \ldots, A_K\}$ such that

$$\beta_{i_1 i_2} \begin{cases} < \frac{\pi}{2}, & \text{if items } i_1 \text{ and } i_2 \text{ come from the same } A_k; \\ > \frac{\pi}{2}, & \text{if items } i_1 \text{ and } i_2 \text{ come from two different } A_k\text{s}; \end{cases} \tag{26}$$

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. $V_{\boldsymbol{\alpha}}^{\perp}$ defined by (12) in Section 2 is a $(d-1)$-dimensional subspace of d-dimensional Euclidean space that is orthogonal to $\boldsymbol{\alpha}$. For example, when $d = 3$, $V_{\boldsymbol{\alpha}}^{\perp}$ is a plane perpendicular to $\boldsymbol{\alpha}$. Since $\dim(V_{\boldsymbol{\alpha}}^{\perp}) = d - 1$, there are at most d vectors in $V_{\boldsymbol{\alpha}}^{\perp}$ such that the angles between any two vectors are greater than $\pi/2$. Therefore, if a d-dimensional test has approximate simple structure, then K, the number of clusters in partition $\mathcal{P}^*$ satisfying (26), will be less than or equal to d, the number of dimensions of the test. This results in the following theorem:

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 (26) 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.

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 structure approximately when their frameworks require that each item simply measure one of several separate subscales. For example, the current grade 4 NAEP reading assessment is assumed to be a two-dimensional simple structure test with each dimension representing one of the two purposes for reading: *reading for literary experience* and *reading to gain information* (see Allen, Donoghue, & Schoeps, 2001). Each item measures only one subscale (one content) and all items measuring the same scale (a one-scale subtest) are considered to be unidimensional. 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, $\mathrm{Cov}(\Theta_i, \Theta_j) > 0$ for all $1 \le i < j \le d$. In fact, they are usually highly correlated. For example, the correlation between the two NAEP reading subscales is approximately 0.85. Obviously, 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 $\mathcal{P}^*$.

Let $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_d)'$ be the weight vector of a composite, and $\boldsymbol{\Sigma} = (\rho_{ij})$ be the correlation matrix of $\boldsymbol{\Theta}$. The *regularity condition* defined by Zhang and Stout (1999b) is said to hold with respect to the composite Θ_α, if

$$\rho_{ij} < \frac{(\sum_{k=1}^{d} \alpha_k \rho_{ik})(\sum_{l=1}^{d} \alpha_l \rho_{jl})}{\sum_{k=1}^{d} \sum_{l=1}^{d} \alpha_k \alpha_l \rho_{kl}} \qquad \text{for all} \quad 1 \le i < j \le d. \tag{27}$$

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. One equivalent form of (27) is that given the composite Θ_α, any two components of $\boldsymbol{\Theta}$ are uncorrelated; that is,

$$\mathrm{Cov}(\Theta_i, \Theta_j \mid \Theta_\alpha) < 0 \qquad \text{for all} \quad 1 \le i < j \le d.$$

In other words, the angles between the projections of Θ_i and Θ_j on the subspace $V_{\boldsymbol{\alpha}}^{\perp}$ defined in Section 2 are all larger than $\pi/2$ for all $1 \le i < j \le d$.

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_js 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_js must be very close (i.e., the respective abilities they load on are highly correlated); these A_js form a large cluster B_k that measures one dominant dimension. In this case, the partition $\mathcal{P}_0 = \{B_1, B_2, \ldots, B_K\}$ satisfies (26). Note that the regularity condition always holds for any two-dimensional simple structure test. To better understand the regularity condition, 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 $\rho_{12} = \rho_{13} = 0.5$, then the regularity condition holds when $\rho_{23} < 0.8028$, but will not hold when $\rho_{23} > 0.8028$ (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 procedure may further be applied to the 40-item math subtest, and it will find the partition with algebra and geometry clusters.

The simple structure assumption is widely used in the analysis of response data. Hence, there is a need for a procedure to check whether a test has simple structure or not, or more generally, whether a test has approximate simple structure or not. Zhang and Stout (1999b) propose 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

$$\mathrm{ASSI}(\mathcal{P}) = \frac{2}{n(n-1)} \sum_{1 \le i_1 < i_2 \le n} \delta_{i_1 i_2}(\mathcal{P}) \mathrm{Sgn}(E[\mathrm{Cov}(X_{i_1}, X_{i_2}|\Theta_T)]) \tag{28}$$

and

$$R(\mathcal{P}) = \frac{\sum_{1 \le i_1 < i_2 \le n} \delta_{i_1 i_2}(\mathcal{P}) E[\mathrm{Cov}(X_{i_1}, X_{i_2}|\Theta_T)]}{\sum_{1 \le i_1 < i_2 \le n} |E[\mathrm{Cov}(X_{i_1}, X_{i_2}|\Theta_T)]|} \tag{29}$$

where Θ_T is a test composite and $\delta_{i_1 i_2}(\mathcal{P})$ is defined by (24). 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.

It is not difficult to show that $\mathrm{ASSI}(\mathcal{P}^*) = 1$ if, and only if, each expected conditional covariance is positive when two items come from the same cluster in $\mathcal{P}^*$ and negative otherwise, according to the definition of $\delta_{i_1 i_2}(\mathcal{P}^*)$. The $\mathrm{ASSI}(\mathcal{P})$ index is decreased whenever a within-cluster item pair produces a negative conditional covariance or a between-cluster item pair produces a positive conditional covariance. Similarly, $R(\mathcal{P}^*) = 1$ if, and only if, each expected conditional covariance is nonnegative when two items come from the same cluster in $\mathcal{P}^*$ and nonpositive otherwise. In fact, all properties of these two indexes for dichotomous items also hold for polytomous items. One can obtain the following theorem using the same proofs from Zhang and Stout (1999b).

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 $\mathrm{ASSI}(\mathcal{P})$ based on conditional covariances is exactly the same as that based on conditional correlation coefficients since by (25)

$$\mathrm{Sgn}(E[\mathrm{Cov}(X_{i_1}, X_{i_2}|\Theta_\alpha)]) = \mathrm{Sgn}(\mathrm{Corr}(X_{i_1}, X_{i_2}|\boldsymbol{\alpha})]).$$

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 $\widehat{\mathrm{Cov}}_{i_1 i_2}$, an estimator of expected conditional covariance, it is easy to construct an estimator of the theoretical DETECT by substituting the expected conditional covariances (or correlations) with their corresponding estimators, that is,

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

where $\delta_{i_1 i_2}(\mathcal{P})$ is defined by (24). Similarly, one can construct estimators for ASSI($\mathcal{P}$) and $R(\mathcal{P})$, that is,

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

and

$$\widehat{R}(\mathcal{P}) = \frac{\sum_{1 \le i_1 < i_2 \le n} \delta_{i_1 i_2}(\mathcal{P}) \, \widehat{\mathrm{Cov}}_{i_1 i_2}}{\sum_{1 \le i_1 < i_2 \le n} \left| \widehat{\mathrm{Cov}}_{i_1 i_2} \right|}.$$

The $\widehat{\mathrm{Cov}}_{i_1 i_2}$ can either be (19) or (21). Similarly, one can use an estimator of conditional correlation $\widehat{\mathrm{Corr}}_{i_1 i_2}$ given by (20) or (22) to construct estimators of DETECT indexes. Currently, the estimators given by (21) and (22) are used in the DETECT program by default.

The operating rule of the DETECT procedure 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 polytomously scored 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, a new 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 $40 \times 1,000$ data set with 40 items and 1,000 examinees is randomly divided into two 40×500 data sets) whenever the number of examinees is large enough. One half-data set is called the target data set, and the other is the reference data set. 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 the DETECT program, 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 target data at $\mathcal{P}_1^*$ and $\mathcal{P}_2^*$. That is, $\widehat{D}_{max} = \widehat{D}_1(\mathcal{P}_1^*)$ and $\widehat{D}_{ref} = \widehat{D}_1(\mathcal{P}_2^*)$. Theoretically, $\widehat{D}_{ref}$ is less than or equal to $\widehat{D}_{max}$. If $\mathcal{P}_1^*$ and $\mathcal{P}_2^*$ are exactly the same, then $\widehat{D}_{ref}$ equals $\widehat{D}_{max}$. If $\widehat{D}_{ref}$ is significantly smaller than $\widehat{D}_{max}$ (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 $\widehat{D}_{Ref}$ 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 $\widehat{\mathrm{ASSI}}_1(\mathcal{P}_2^*)$ or $\widehat{\mathrm{R}}_1(\mathcal{P}_2^*)$ is larger than a critical value will the DETECT program 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 $\widehat{\mathrm{ASSI}}_1(\mathcal{P}_2^*)$ and $\widehat{\mathrm{R}}_1(\mathcal{P}_2^*)$ 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, DETECT will ask its user to use other information to determine the dimensional structure of response data.

As discussed by Zhang and Stout (1999b), when the DETECT program 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. 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 is $n/13$ truncated to be between 3 and 9, where n is the number

of items in the test.

The DETECT program 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.

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. For instance, the fourth grade NAEP reading assessment is assumed to be a two-dimensional simple structure with each dimension representing one of the two purposes for reading. It is important to know whether or not the statistical dimensional structure is in concert with the above substantive test structure, or more practically to check if the content-based partition is near optimal. 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 DETECT program lets its user provide such a partition for a confirmatory analysis.

Various types of errors can happen in the applications of DETECT. One type of possible error is that DETECT fails to detect the unidimensionality of response data, and another is that it misjudges a multidimensional response data set as unidimensional. The other two types of possible errors are that DETECT incorrectly enumerates the number of dimensions and/or assigns some items to wrong clusters. It can be expected that the rates of those errors should be low when the numbers of items and examinees are large and may be high otherwise. In Section 6, simulation studies will be conducted to check the rates of errors or the correct rates for various cases.

5. Applying DETECT to NAEP Reading Data

In this section, the DETECT procedure is used to analyze the dimensional structure of the 2002 NAEP reading grades 4 and 8 operational data.

According to the contemporary definition of reading literacy, the NAEP reading items

were developed in accordance with three contexts or purposes for reading and four aspects of reading or reading stances (see National Assessment Governing Board, 1992). The three contexts or purposes for reading are *reading for literary experience*, *reading to gain information*, and *reading to perform a task*. These three reading contexts are assessed across four reading aspects (stances) that include *forming an initial understanding*, *developing an interpretation*, *personal reflection and response*, and *demonstrating a critical stance*.

NAEP assesses all three contexts for reading in grade 8, but only the first two contexts in grade 4. The proportion of items related to each context for reading changes from grade to grade to reflect the changing demands made of students as they mature. The target percentage of items related to each context for reading or each aspect of reading is specified in the NAEP Reading Framework.

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 numbers of multiple-choice and dichotomously and polytomously scored constructed-response items in the 2002 NAEP reading assessment were 37, 27, and 18, respectively, for grade 4, and 42, 38, and 31, respectively for grade 8. The test composition is listed in Table 1.

Because of the test time limitation, no one student takes all the items. Reading passages and accompanying items are divided into blocks in all main NAEP reading 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). 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). Hence, each sampled student took

about 20 items, which leads to a lot of missing data by this design. Missing data are treated according to the NAEP conventions in this study. Missing responses at the end of a block of items are considered not reached and are treated as if they had not been presented to the student. Missing responses to items before the last observed response in a block are considered intentional omissions. If the omitted item is a multiple-choice item, the missing response is treated as fractionally correct at the value of the reciprocal of the number of response alternatives. If the omitted item is not a multiple-choice item, the missing response is scored zero.

Table 1

2002 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 is for multiple-choice items, and CR-D and CR-P are for constructed-response items scored dichotomously and polytomously, respectively.

The reporting scales of NAEP reading assessments are based on the contexts for reading in the reading operational analysis conducted at ETS. It has been assumed that the NAEP reading assessments of grades 4 and 8 have two and three subscales, respectively, and that each subscale represents one context (purpose) for reading. 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. 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. Plausible values are drawn from the joint distribution of scale proficiency values for each assessed student. 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 the last section, 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 well reflects 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 DETECT program with cross validation. Recall that the DETECT 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, this paper only presents the results based on conditional covariances here.

The DETECT results show that the fourth-grade data set is two dimensional. The optimal two-cluster partition provided by DETECT is consistent with the two contexts for reading, 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, which is certainly not a good item from the psychometric point of view (low item information). Note that this item will be excluded in our simulation study in the next section.

DETECT concludes that the eighth grade response data set is three-dimensional; its optimal three-cluster partition agrees with the three contexts for reading, except for item 40 and the 50-minute block. 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 the case that no students complete a pair of items, the DETECT program automatically assigns 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, DETECT always keeps all the 50-minute block items in the same cluster, indicating that these items are relatively dimensionally homogeneous given the composite.

Table 2 presents the three index values at the three partitions from the DETECT 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 2 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 weak 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 2

DETECT Results for the 2002 NAEP Reading Data for Grades 4 and 8

Partition	Grade 4			Grade 8		
	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

6. Simulation Studies

In this section, simulation studies were conducted to check the performance of DETECT based on the estimators given by (21) and (22), which are the DETECT default estimators. In the simulation studies, response data sets were generated to be either unidimensional or multidimensional with simple structure. Then DETECT 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. The previous section showed that there is a bad item in the second subscale for each grade.

These two items were excluded from these simulation studies. Therefore, the total number of items for grade 4 in the simulation is 81, where 41 items measure the first subscale, and 40 items measure the second subscale; while for grade 8, the number of items used is 110 with 30, 47, and 33 items measuring each of the three reading subscales, 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 its length less than 81 in grade 4 cases, or less than 110 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 DETECT run with cross validation is 1,000, 2,000, 4,000, 6,000, 8,000, or 10,000, with half as a target data set and 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 1, subscales are the same and corresponding 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 DETECT to identify its dimensional structure) was replicated 100 times.

The results are summarized in Table 3 through Table 6. Table 3 and Table 5 present the counts out of 100 replications that DETECT correctly declared the number of dimension(s) for grades 4 and 8, respectively. Table 4 and Table 6 list the counts out of 100 replications for which DETECT not only correctly declared the number of dimension(s) but also correctly classified items into dimensionally based clusters; that is, the optimal partition of items found by DETECT is exactly the true dimensionally based one. There are two numbers in each cell of these four tables: The first is the count obtained by DETECT based on conditional covariances, and the second is based on conditional correlations. Note that if the theoretical DETECT were used, all numbers in these four tables should be 100.

Table 3

Frequency of Correct DETECT Results out of 100 Replications Using Conditional Covariance/Correlation in Unidimensional or Two-Dimensional Cases (Grade 4)

n	ρ	Number of examinees					
		1,000	2,000	4,000	6,000	8,000	10,000
	0.0	100/100	100/100	100/100	100/100	100/100	100/100
	0.3	100/100	100/100	100/100	100/100	100/100	100/100
	0.6	100/100	100/100	100/100	100/100	100/100	100/100
20	0.8	100/99	100/100	100/100	100/100	100/100	100/100
	0.9	22/18	79/81	100/100	100/100	100/100	100/100
	1.0	100/100	100/100	95/90	81/77	58/46	41/26
	0.0	100/100	100/100	100/100	100/100	100/100	100/100
	0.3	100/100	100/100	100/100	100/100	100/100	100/100
	0.6	100/100	100/100	100/100	100/100	100/100	100/100
40	0.8	98/97	100/100	100/100	100/100	100/100	100/100
	0.9	11/11	91/94	100/100	100/100	100/100	100/100
	1.0	100/100	100/100	100/100	100/100	100/100	100/100
	0.0	100/100	100/100	100/100	100/100	100/100	100/100
	0.3	100/100	100/100	100/100	100/100	100/100	100/100
	0.6	100/100	100/100	100/100	100/100	100/100	100/100
60	0.8	100/100	100/100	100/100	100/100	100/100	100/100
	0.9	14/12	94/94	100/100	100/100	100/100	100/100
	1.0	100/100	100/100	100/100	100/100	100/100	100/100
	0.0	100/100	100/100	100/100	100/100	100/100	100/100
	0.3	100/100	100/100	100/100	100/100	100/100	100/100
	0.6	100/100	100/100	100/100	100/100	100/100	100/100
81	0.8	100/100	100/100	100/100	100/100	100/100	100/100
	0.9	14/9	99/99	100/100	100/100	100/100	100/100
	1.0	100/100	100/100	100/100	100/100	100/100	100/100

Note. n is the number of items. ρ is the population correlation coefficient between two subscales. $\rho = 1.0$ represents a unidimensional case.

Table 4

Frequency of Completely Correct DETECT Results out of 100 Replications Using Conditional Covariance/Correlation in Unidimensional or Two-Dimensional Cases (Grade 4)

n	ρ	Number of examinees					
		1,000	2,000	4,000	6,000	8,000	10,000
	0.0	100/100	100/100	100/100	100/100	100/100	100/100
	0.3	100/100	100/100	100/100	100/100	100/100	100/100
	0.6	100/99	100/100	100/100	100/100	100/100	100/100
20	0.8	72/68	94/92	100/100	100/100	100/100	100/100
	0.9	6/9	35/41	84/85	88/91	95/94	99/97
	1.0	100/100	100/100	95/90	81/77	58/46	41/26
	0.0	100/100	100/100	100/100	100/100	100/100	100/100
	0.3	100/100	100/100	100/100	100/100	100/100	100/100
	0.6	98/97	100/100	100/100	100/100	100/100	100/100
40	0.8	54/53	84/84	96/96	99/98	100/99	100/100
	0.9	3/3	25/27	63/58	71/78	81/77	86/85
	1.0	100/100	100/100	100/100	100/100	100/100	100/100
	0.0	100/100	100/100	100/100	100/100	100/100	100/100
	0.3	100/100	100/100	100/100	100/100	100/100	100/100
	0.6	95/95	100/100	100/100	100/100	100/100	100/100
60	0.8	46/44	88/87	99/99	99/100	100/100	100/100
	0.9	4/3	26/23	66/65	85/87	87/89	93/94
	1.0	100/100	100/100	100/100	100/100	100/100	100/100
	0.0	100/100	100/100	100/100	100/100	100/100	100/100
	0.3	100/100	100/100	100/100	100/100	100/100	100/100
	0.6	94/94	99/99	100/100	100/100	100/100	100/100
81	0.8	56/54	90/92	100/100	100/100	100/100	100/100
	0.9	2/1	22/25	80/79	92/92	97/97	100/100
	1.0	100/100	100/100	100/100	100/100	100/100	100/100

Note. n is the number of items. ρ is the population correlation coefficient between two subscales. $\rho = 1.0$ represents a unidimensional case.

Table 5

Frequency of Correct DETECT Results out of 100 Replications Using Conditional Covariance/Correlation in Unidimensional or Three-Dimensional Cases (Grade 8)

n	ρ	Number of examinees					
		1,000	2,000	4,000	6,000	8,000	10,000
45	0.0	100/100	100/100	100/100	100/100	100/100	100/100
	0.3	100/100	100/100	100/100	100/100	100/100	100/100
	0.6	100/100	100/100	100/100	100/100	100/100	100/100
	0.8	96/93	100/100	100/100	100/100	100/100	100/100
	0.9	0/0	30/12	99/100	1100/100	100/100	100/100
	1.0	100/100	100/100	100/100	100/100	100/100	100/100
60	0.0	100/100	100/100	100/100	100/100	100/100	100/100
	0.3	100/100	100/100	100/100	100/100	100/100	100/100
	0.6	100/100	100/100	100/100	100/100	100/100	100/100
	0.8	99/97	100/100	100/100	100/100	100/100	100/100
	0.9	0/0	37/25	100/100	100/100	100/100	100/100
	1.0	100/100	100/100	100/100	100/100	100/100	100/100
90	0.0	100/100	100/100	100/100	100/100	100/100	100/100
	0.3	100/100	100/100	100/100	100/100	100/100	100/100
	0.6	100/100	100/100	100/100	100/100	100/100	100/100
	0.8	100/100	100/100	100/100	100/100	100/100	100/100
	0.9	0/0	73/47	100/100	100/100	100/100	100/100
	1.0	100/100	100/100	100/100	100/100	100/100	100/100
110	0.0	100/100	100/100	100/100	100/100	100/100	100/100
	0.6	100/100	100/100	100/100	100/100	100/100	100/100
	0.8	100/100	100/100	100/100	100/100	100/100	100/100
	0.9	0/0	50/21	100/100	100/100	100/100	100/100
	1.0	100/100	100/100	100/100	100/100	100/100	100/100

Note. n is the number of items. ρ is the population correlation coefficient between two subscales. $\rho = 1.0$ represents a unidimensional case.

Table 6

Frequency of Completely Correct DETECT Results out of 100 Replications Using Conditional Covariance/Correlation in Unidimensional or Three-Dimensional Cases (Grade 8)

n	ρ	Number of examinees					
		1,000	2,000	4,000	6,000	8,000	10,000
45	0.0	99/99	100/100	100/100	100/100	100/100	100/100
	0.3	95/95	100/100	100/100	100/100	100/100	100/100
	0.6	76/76	97/95	99/99	100/100	100/100	100/100
	0.8	15/18	51/55	90/88	93/94	97/97	96/96
	0.9	0/0	2/1	19/17	35/31	39/44	44/53
	1.0	100/100	100/100	100/100	100/100	100/100	100/100
60	0.0	100/100	100/100	100/100	100/100	100/100	100/100
	0.3	98/97	100/100	100/100	100/100	100/100	100/100
	0.6	81/79	99/98	100/100	100/100	100/100	100/100
	0.8	9/6	53/52	93/93	94/95	98/99	99/99
	0.9	0/0	3/1	28/24	41/37	50/53	62/71
	1.0	100/100	100/100	100/100	100/100	100/100	100/100
90	0.0	100/100	100/100	100/100	100/100	100/100	100/100
	0.3	95/96	100/100	100/100	100/100	100/100	100/100
	0.6	76/79	99/99	100/100	100/100	100/100	100/100
	0.8	11/6	69/64	97/95	98/98	100/100	100/100
	0.9	0/0	4/4	37/41	74/73	80/80	88/89
	1.0	100/100	100/100	100/100	100/100	100/100	100/100
110	0.0	100/100	100/100	100/100	100/100	100/100	100/100
	0.3	96/95	100/100	100/100	100/100	100/100	100/100
	0.6	80/76	99/99	100/100	100/100	100/100	100/100
	0.8	10/10	56/55	98/98	98/99	99/99	100/100
	0.9	0/0	2/4	38/39	81/78	89/87	97/96
	1.0	100/100	100/100	100/100	100/100	100/100	100/100

Note. n is the number of items. ρ is the population correlation coefficient between two subscales. $\rho = 1.0$ represents a unidimensional case.

Tables 4 and 6 show that DETECT found the true dimensionally based partition in every replication when the correlation was low or moderate and the number of examinees was large. DETECT 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. DETECT successfully identified the unidimensional cases in every replication in every case except for the cases of 20 items and 4,000 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, a conditional covariance based procedure 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 1,000, DETECT does not seem to perform well: Most times DETECT 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 1,000, and the correlation coefficient is 0.9, the frequency with which DETECT 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 3). 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 3 and 5 that DETECT correctly declared the number of dimension(s) are 96% and 95% for grades 4 and 8, respectively.

The results in these four tables also indicate that the performance of DETECT based on conditional covariances is almost the same as that based on conditional correlations in the situations considered in this simulation study. Because of length limitations of a paper, I do not report all DETECT statistics here. However, 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 4,000 or more examinees. The maximum DETECT value has a negative association 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 the test displays, and the correlation of the underlying abilities (subscales) is one of the important factors in determining the degree of multidimensionality.

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 new 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 randomly. DETECT 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 DETECT based on its default estimates of conditional covariances/correlations defined in Section 3 performs pretty well.

In this paper, DETECT 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, DETECT indicates that the 2002 NAEP reading data sets are weakly multidimensional. That is, it is still not unreasonable to treat the data sets as essentially unidimensional. 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. This suggests that it is reasonable to analyze the NAEP reading data based on the contexts/purposes for reading, a substantive test structure that is currently adopted in the NAEP reading operational analysis. Yu and Nandakumar (2001) apply 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.

Considering that there is no fixed procedure for assessing dimensionality structure in the process of NAEP operational analysis, DETECT can be an excellent dimensionality assessment tool in the post-hoc dimensionality analysis for checking whether the context-based or content-based classification of items into subscales is statistically consistent with the dimensional structure of the data.

Although the new version of DETECT works well, it is still an open question as to how to construct a consistent unbiased estimator for $E[\mathrm{Cov}(X_i, X_j|\Theta_T)]$ 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, U.S. 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, U.S. Department of Education.

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

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, April). *Exploring and explaining the lack of local independence through conditional covariance functions.* Paper presented at the 1994 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.

Johnson, R. A., & Wichern, D. W. (1992). *Applied multivariate statistical analysis*. Upper Saddle River, NJ: Prentice Hall.

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, Canada: University of Ottawa Press.

Mislevy, R., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics, 17*, 131-154.

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

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.

Rosenbaum, P. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika, 49*, 425-435.

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

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

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, April). *Construction and evaluation of bias-corrected estimators of DETECT dimensionality index.* Paper presented at the annual meeting of 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.