Chapter 5

TEST MODELING

RATNA NANDAKUMAR

TERRY ACKERMAN

Discoveries with item response theory (IRT) principles, since the 1960s, have led to major breakthroughs in psychological and educational assessment. For example, using IRT principles, it is possible to determine the relative standing of an examinee on the latent continuum by administering any sample of items from a given domain of knowledge. This is possible through the principle of invariance in IRT, which means that item properties such as difficulty and discrimination can be determined irrespective of the ability level of the examinee. Hence, any set of items from a given domain can be used to estimate an examinee's position along the latent continuum. This is in sharp contrast to the traditional classical test theory (CTT), in which item statistics are a function of the specific group of examinees who took the item, and the examinee's performance is a function of the items on the test. That is, in CTT, the same item may have different p-values depending on the level of the examinees' ability taking the item. Similarly, in CTT, it is not possible to generalize the performance of an examinee beyond a given set of test items.

The advantages of IRT techniques are associated with strong models used to characterize examinee performance on a test, as opposed to the weak models of CTT that are tautologies and not testable. One can realize the potentials of IRT modeling and its consequences only if there is a close match between the model and data. Application of IRT techniques to data without ensuring the model-data fit can lead to unfair and unjustified ranking of examinees on the latent continuum of domain of interest.

The fundamental underlying assumptions of item response models are monotonicity, dimensionality, and local independence. Monotonicity implies that item performance is monotonically related to the ability. That is, a high-ability examinee has a greater probability of responding correctly to the item than a low-ability examinee. Because achievement test items inherently satisfy this assumption, it is implicitly assumed.[1] Local independence (LI) implies that item responses are conditionally independent. The conditional ability vector that ensures item independence is key to determining the dimensionality of data. For example, if local independence is achieved by conditioning on a unidimensional latent trait, then the response data are said to be unidimensional. If local independence is achieved by conditioning on a two-dimensional latent trait vector, then the response data are said to be two-dimensional. Hence, local independence and dimensionality assumptions are intertwined. One can only statistically test either of the assumptions assuming the other.

In addition to these basic foundational assumptions, a given model may have other assumptions. For

1. Normally, during the test construction process, if an item does not satisfy the assumption of monotonicity, it is deleted from the test.

example, among parametric models, there are models associated with different item types, such as dichotomous items (item is scored correct vs. incorrect) and polytomous items (arising from scoring essays and performance-type tasks). Each model has a set of assumptions associated with it. For a list of IRT models for different item formats and their development, refer to van der Linden and Hambleton (1997). To date, a great majority of tests are intended to be unidimensional ($d = 1$). That is, the purpose of the test is to assess an examinee's trait level based on his or her responses to unidimensional test items. Examinee test performance on a unidimensional test can be summarized with a single scale score. It is also well known that any unidimensional test is typically influenced by transient dimensions (abilities) common to just a few of the items. It is well documented (Hambleton & Swaminathan, 1985; Humphreys, 1985, 1986; Stout, 1987) that summarizing examinees' performance with a single scale score in the presence of transient abilities is harmless. However, when transient abilities are not insignificant, such as a paragraph comprehension test, or when a test is intentionally multidimensional, then a single scale score is not a meaningful format to summarize examinee performance. A multidimensional or other appropriate model is needed to summarize examinee performance. Hence, given test data, we need to empirically determine if unidimensional modeling and the resulting single-scale score summary is meaningful. If unidimensional modeling is not appropriate, ways to go about selecting an appropriate model are needed.

The focus of this chapter is to illustrate modeling of dichotomous data. Both unidimensional and multidimensional modeling are considered. In the following sections, assumptions of local independence and dimensionality are defined; several tools for assessing these assumptions will be described, and these tools will be illustrated with several realistic data sets. Based on these tools and indices, guidelines for determining an appropriate model for given data will be delineated.

5.1. Definition of Local Independence and Dimensionality

The purpose of a majority of standardized tests is to measure a single construct, ability, or dimension. Hence, a major question facing any test development, analysis, and interpretation is whether it is appropriate to summarize the performance of an examinee to test items using a single scaled score. That is, can the test be modeled using a monotone, locally independent, unidimensional model? The answer is simple. If the test items are tapping a single construct or one dominant dimension, and if the examinee subpopulation taking the test is homogeneous with respect to the construct being measured, then a single scaled score will summarize examinees' performance on the test. Although the answer is simple, ways of determining that the test indeed is measuring a dominant construct is not so simple. Assuming that the assumption of monotonicity is checked and satisfied during the test development process,[2] let us examine the definitions of local independence and dimensionality.

Let $\mathbf{U_n} = (U_1, U_2, \ldots, U_n)$ denote the item response pattern of a randomly sampled examinee on a test of length n. The random variable U_i takes a value of 1 if the item is correctly answered and 0 if the item is incorrectly answered. Let Θ denote the latent ability, possibly multidimensional, underlying item responses.

Definition 1. The test items $\mathbf{U_n}$ are said to be *locally independent* if

$$\text{Prob}(\mathbf{U_n} = \mathbf{u_n}|\boldsymbol{\Theta} = \boldsymbol{\theta}) = \prod_{i=1}^{n} \text{Prob}(U_i = u_i|\boldsymbol{\theta}) \quad (1)$$

for each response pattern $\mathbf{u_n} = (u_1, u_2, \ldots, u_n)$ and for all $\boldsymbol{\theta}$. That is, conditional on examinee ability, responses to different items are independent.

The dimensionality d of a test $\mathbf{U_n}$ is the minimal dimensionality required for $\boldsymbol{\Theta}$ to produce a model that is both monotone and locally independent (Stout, Habing, Douglas, Kim, Roussos, & Zhang, 1996). When $\boldsymbol{\Theta}$ consists of a single component, θ, then the test is said to be unidimensional. The definition of local independence provided above is referred to as the strong local independence (SLI) as it involves complete independence among items conditioned on examinee ability. On the other hand, weak local independence (WLI) involves conditional item pair covariance to be zero for all items pairs. That is, $\text{cov}(U_i, U_j|\theta) = 0$.

Definition 2. The test items $\mathbf{U_n}$ are said to be *weakly locally independent* if

$$\text{Prob}(U_i = u_i, U_j = u_j|\Theta = \theta) = \text{Prob}(U_i = u_i|\Theta = \theta)\text{Prob}(U_j = u_j|\Theta = \theta) \quad (2)$$

for all $n(n-1)/2$ items pairs and for all θ. WLI is also referred to as pairwise local independence (McDonald, 1994, 1997). Obviously, SLI implies WLI. It is

2. Monotonicity of items is established by high positive point-biserial correlation between the item score and the test score.

commonly accepted that, if the unidimensionality can be achieved through pairwise local independence, then unidimensionality is closely approximated through SLI (Stout, 2002).

From a factor-analytic point of view, it is not realistic to construct a strictly unidimensional test. In any test, it is not uncommon to find transient abilities common to one or more items (Humphreys, 1985; Tucker, Koopman, & Linn, 1969). In this sense, unidimensionality refers to the dominant ability measured by the test. The WLI, although very useful for empirical investigation of a dimensional structure underlying test data, does not capture the concept of dominant dimensions underlying data.

Stout (1987, 1990, 2002) theoretically conceptualized the separation of dominant dimensions from inessential or transient dimensions and referred to them as *essential dimensions,* meaning what the test is essentially measuring. Stout (1987) also developed a statistical test of essential unidimensionality. In his conceptual formulation and definition of essential dimensionality, Stout (1990) used the "infinite-length test" abstraction. That is, to understand the structure underlying test data resulting from administering a finite test to a finite group of examinees, Stout derived theoretical foundational results based on the abstraction of an infinite-length test U_∞ administered to a large group of examinees. Using this conceptual framework of infinite-length test, essential dimensionality is defined as follows.

Definition 3. A test U_∞ is essentially unidimensional with respect to the unidimensional latent random variable Θ if, for all θ,

$$\frac{\sum_{1\le i<j\le n} |\mathrm{Cov}(U_i, U_j|\Theta=\theta)|}{\binom{n}{2}} \to 0, \qquad (3)$$

as $n \to \infty$. The above definition implies that the average covariance, in the limit, approaches 0 as the test length increases to ∞. In other words, transient or nonessential traits common to one or more items may result in nonzero conditional covariance. However, the average covariance approaches 0. Essential dimensionality is a weaker form of strict dimensionality based on either SLI or WLI.

The definition of essential dimensionality has further led to theoretical results establishing the usefulness of number-correct score as a consistent estimator of unidimensional ability on the latent true-score scale (Stout, 1990) and to nonparametric estimation of item response functions (Douglas & Cohen, 2001).

5.2. Geometrical Representation of Multidimensional Structure

Although, in reality, dimensionality is determined by test items together with the examinee population taking the test, the geometrical description of items in the latent space provides an intuitive understanding of how item *direction* with respect to the test *direction* contributes to the dimensional structure underlying test data. In explaining the dimensional structure of test items geometrically, only two-dimensional test items are considered.

An item can be geometrically represented by a vector, which, if extended, passes through the origin of a coordinate system. The coordinate axes represent the two dimensions, θ_1 and θ_2, underlying test data. The origin of the coordinate system is the population multidimensional trait-level mean. The direction of the vector represents the θ_1, θ_{2-}, composite that has the maximum discrimination, which is appropriately defined for the model in use. The length of the vector is a measure of the magnitude of the item's discrimination, denoted by MDISC $= (a_1^2 + a_2^2)^{1/2}$, where a_1 and a_2 are the discriminating parameters associated with the two dimensions. The location of the base of the item vector corresponds to that level of multidimensional ability at which the probability of correct response to the item is 0.5. The item vector is orthogonal to the $p = .5$ equiprobability contour (Ackerman, 1996; Reckase, 1997). For example, in a two-dimensional space, items are located only in the first or third quadrants. This is because item discriminations can only take positive values. Easy items are located in the third quadrant and difficult items in the first quadrant. Figure 5.1 shows vector representation of items in a two-dimensional space. Item 1 is an easy item with low discrimination, whereas Item 2 and Item 3 are more difficult and high-discriminating items. The angle direction of the item measured from the θ_1-axis represents a composite of dimensions that the item is best measuring. For example, in a two-dimensional space, if the angle distance of an item from the θ_1-dimension is small, then the item is measuring mostly the θ_1-dimension (Item 3 in Figure 5.1). On the other hand, if the item vector is at 45 degrees, then the item's ability composite measures both dimensions equally (Item 1 in Figure 5.1).

Intuitively speaking, a test of items whose vectors cluster in a narrow sector (i.e., where all the items are measuring similar ability composites) is considered to be essentially unidimensional. If all test items lie on the coordinate axis (as opposed to a narrow sector), then the test would be considered strictly unidimensional.

Figure 5.1 Vector Representation of Two-Dimensional Items

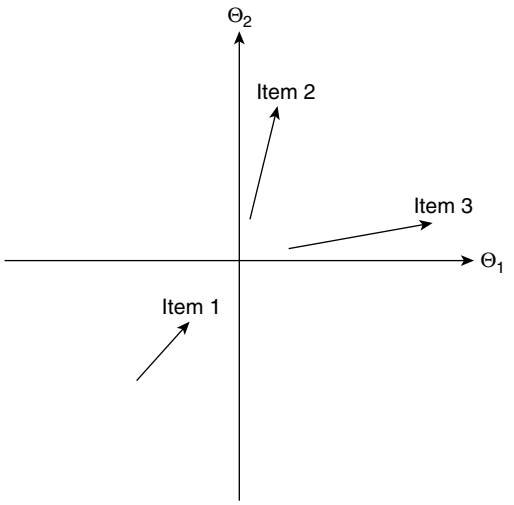
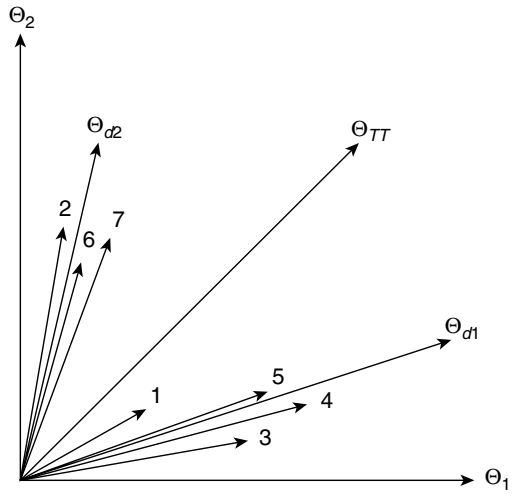


Figure 5.2 An Example of an Approximate Simple Structure Test



The way item vectors cluster together with respect to the coordinate axes, in a multidimensional space, determines the dimensional structure of the test. For the two-dimensional latent space, Figure 5.2 provides an example of a test with two clusters, whose direction of best measurement is represented by vectors Θ_{d1} and Θ_{d2}.[3] The direction of best-measurement vector Θ_{d1} is a weighted average of item discrimination vectors comprising its cluster. The same is true for Θ_{d2}. The direction of best measurement of the total test comprising the two clusters is represented by the vector Θ_{TT}.

A test is considered to have *simple structure* if all items in the test lie along the coordinate axes. In this case, although the dimensional clusters may be correlated, each is an independent item cluster. If, on the other hand, test items are spread along a narrow sector surrounding the coordinate axes, then each narrow sector of items is considered exhibiting an *approximate simple structure.* Figure 5.2 illustrates an example of an approximate simple structures test with two item clusters. Mathematically speaking, *approximate simple structure* can be defined as a k-dimensional latent coordinate axis existing within a d-dimensional latent space ($d \geq k$) such that items only lie within narrow sectors surrounding the coordinate axis. In such a case, there are k-dominant dimensions (Stout et al., 1996).

Zhang and Stout (1999a) have proved theoretical results for using conditional covariances as the basis for determining the dimensional structure underlying multidimensional data. The central theme of their results is that the dimensional structure of test data can be completely discovered using item pair conditional covariances (CCOV), conditional on the test vector represented by Θ_{TT}, provided there is an approximate simple structure underlying test data. The pattern of CCOV_{ij} is positive if items i and j measure similar ability composites, negative if items i and j measure different ability composites, and 0 if one of the items measures the same composite as Θ_{TT}. For example, in the case of a two-dimensional structure, as in Figure 5.2, the CCOV of an item pair is positive if the item vectors in the pair lie on the same side of the conditioning variable's direction of best measurement, Θ_{TT} (e.g., Items 3 and 4). The CCOV is negative if the item vectors lie on the opposite sides of Θ_{TT} (e.g., Items 1 and 2). The CCOV is zero if one of the items lies near the direction of best measurement, Θ_{TT}. This reasoning has been generalized to higher dimensions by Zhang and Stout through $d-1$ dimensional hyperplanes orthogonal to Θ_{TT} and by projecting each item onto this hyperplane.

The magnitude of CCOV indicates the degree of closeness of items' directions of best measurement to each other and their closeness to the conditional vector, Θ_{TT}. CCOV increases as the angle between item pair vectors decreases and as the angle either of the items makes with the Θ_{TT}-axis increases. The CCOV also relates to the degree of discrimination of the vectors. The CCOV increases in proportion to the items' discrimination vectors. Hence, CCOVs form

3. Coordinate axes are not necessarily orthogonal. For example, if $\text{cov}(\Theta_i, \Theta_j) > 0$, then the coordinate axes are not orthogonal.

the basis for establishing the dimensional structure underlying given data. Methods for assessing dimensional structure based on CCOVs are described and illustrated below.

5.3. Methods to Assess the Dimensional Structure Underlying Test Data

This section describes nonparametric methodologies for empirically determining the dimensional structure underlying test data based on CCOVs. It is assumed that one would use these procedures after the test is well developed and its reliability and validity have been established. As explained earlier, it is very important to assess the dimensional structure of the test to determine the test scoring and related issues such as equating and differential item functioning. If the unidimensional model is not appropriate, then recommendations will be made about finding an appropriate model.

Nonparametric tools DIMTEST and DETECT will be used to illustrate the steps involved in determining the correct model for data. We chose these methods because they are not dependent on any particular parametric model for scoring and describing data, and they are simple and easy to use. DIMTEST and DETECT are described below, followed by a flowchart to correctly determine the appropriate model for given data.

5.3.1. DIMTEST

DIMTEST (Stout, 1987; Nandakumar & Stout, 1993; Stout, Froelich, & Gao, 2001) is a nonparametric statistical procedure designed to test the hypothesis that the test data were generated from an LI, $d = 1$ model. The procedure for testing the null hypothesis consists of two steps. In Step 1, n test items are partitioned into two subtests, AT and PT. The AT subtest is of length $m (4 \leq m <$ half the test length), and the PT subtest is of length $n - m$. The AT subtest consists of items that are believed to be dimensionally homogeneous, and the PT subtest consists of the remaining items of the test. One way to select items for AT and PT subtests is through linear factor analysis of the tetrachoric correlation matrix (Froelich, 2000; Hattie, Krakowski, Rogers, & Swaminathan, 1996; Nandakumar & Stout, 1993). This is an automated procedure that uses part of the sample to select items for AT and PT subtests. Items loading on the same dimension are selected into the AT subtest. Expert opinion is another way to select items into these subtests (Seraphine, 2000). Because of the manner in which items are selected, when multidimensionality is present in test data, items in the AT subtest will be predominantly measuring the same unidimensional construct, whereas the remaining items in the PT subtest will be multidimensional in nature. If, on the other hand, the test is essentially unidimensional, then items in both the AT and PT subtests will be measuring the same valid unidimensional construct.

In Step 2, the DIMTEST statistic, T, is computed as follows. Examinees are grouped into subgroups based on their score on the PT subtest consisting of $n - m$ items. The kth subgroup consists of examinees whose total score on the PT subtest, denoted by X_{PT}, is k. In each subgroup k, two variance components, $\hat{\sigma}_k^2$ and $\hat{\sigma}_{U,k}^2$, are computed using items in the AT subtest:

$$\hat{\sigma}_k^2 = \frac{1}{J_k} \sum_{j=1}^{J_k} (Y_j^{(k)} - \bar{Y}^{(k)})^2$$

and

$$\hat{\sigma}_{U,k}^2 = \sum_{i=1}^{m} \hat{P}_i^{(k)} (1 - \hat{p}_i^{(k)}),$$

where

$$Y_j^{(k)} = \sum_{i=1}^{m} U_{ij}^{(k)}, \quad \bar{Y}^{(k)} = \frac{1}{J_k} \sum_{j=1}^{J_k} Y_j^{(k)},$$

$$\hat{p}_i^{(k)} = \frac{1}{J_k} \sum_{j=1}^{J_k} U_{ij}^{(k)},$$

and U_{ij}^k denotes the response of the jth examinee from subgroup k to the ith assessment item in AT, and J_k denotes the number of examinees in the subgroup k. After eliminating sparse subgroups containing too few examinees, let K denote the total number of subgroups used in the computation of the statistic T.

For each examinee subgroup k, compute

$$T_{L,k} = \hat{\sigma}_k^2 - \hat{\sigma}_{U,k}^2 = 2 \sum_{i < l \in AT} \widehat{\mathrm{Cov}}(U_i, U_l | X_{PT} = k),$$

where $\widehat{\mathrm{Cov}}(U_i, U_l | X_{PT} = k)$ is an estimate of the covariance between items U_i and U_l for examinees whose score on the PT subtest is k.

The statistic T_L is given by

$$T_L = \frac{\sum_{k=1}^{K} T_{L,k}}{\sqrt{\sum_{k=1}^{K} S_k^2}},$$

where S_k^2 is the appropriately computed asymptotic variance (Nandakumar & Stout, 1993; Stout et al., 2001) of the statistic $T_{L,k}$. For finite test lengths, the statistic T_L is known to exhibit positive bias (Stout, 1987). The positive bias in T_L is eliminated using a bootstrap technique as follows: For each item, an estimate of its unidimensional item response function (IRF) is computed using a kernel-smoothing procedure (Douglas, 1997; Ramsay, 1991). Using the estimated IRFs, examinee responses are generated for each of the items. Using the generated data and the original AT and PT subtest partition, another DIMTEST statistic is computed, denoted by T_G (see Froelich, 2000, for details). This process of the random generation of unidimensional data with kernel-smoothed estimates of items and the computation of T_G is repeated N times, and the average is denoted by $\bar{T}_G$. $\bar{T}_G$ denotes the inflation or bias in T_L that is due to the finite test length administered to a finite sample of examinees. The final bias-corrected DIMTEST statistic T is given by

$$T = \frac{T_L - \bar{T}_G}{\sqrt{(1 + 1/N)}}. \tag{4}$$

The statistic T follows the standard normal distribution as the number of items and the number of examinees tend to infinity. The null hypothesis of unidimensionality is rejected at level α if T is larger than the $100(1-\alpha)$th percentile of the standard normal distribution.

A number of studies have found the DIMTEST to be a reliable and consistent methodology for assessing unidimensionality. It is also extremely powerful compared to other methodologies in its power to detect multidimensionality (Hattie et al., 1996; Nandakumar, 1993, 1994; Nandakumar & Stout, 1993). The current version of DIMTEST, with recent revisions by Stout et al. (2001), is even more powerful than the former version and can be applied on test sizes as small as 15 items.

5.3.2. DETECT

DETECT (Kim, 1994; Zhang & Stout, 1999a, 1999b) is a statistical methodology for determining the multidimensional structure underlying test data. It partitions the test items into clusters in such a manner that items within clusters are dimensionally cohesive. The DETECT methodology uses the theory of conditional covariances to arrive at the partitioning of test items into clusters. As a result, items within a cluster have positive CCOVs with each other; and items from different clusters have negative CCOVs. The DETECT procedure also quantifies the degree of multidimensionality present in given test data. It is important to note that the number of dimensions and the degree of multidimensionality are two distinct pieces of information. For example, one could have a two-dimensional test in which the two item clusters are dimensionally far apart or close together. In the former case, the degree of multidimensionality is more than in the latter case. For example, in Figure 5.2, clusters represented by vectors Θ_{d1} and Θ_{d2} are the two dimensions underlying test data comprising all test items. The angle between these two vectors determines the degree of multidimensionality present in test data. If the angle between vectors Θ_{d1} and Θ_{d2} is small, the degree of multidimensionality present in test data is small, implying that the two clusters are dimensionally similar. If, on the other hand, the angle between the vectors is large, then two item clusters are dimensionally apart.

The theoretical computation of the DETECT index is briefly described here (for details, see Zhang & Stout, 1999b). Let n denote the number of dichotomous items of a test. Let $P = \{A_1, A_2, \ldots, A_k\}$ denote a partition of the n test items into k clusters. The theoretical DETECT index $D(P)$, which gives the degree of multidimensionality of the partition P, is defined as

$$D(P) = \frac{2}{n(n-1)} \times \sum_{1 \le i < j \le N} \delta_{ij} E[\mathrm{Cov}(X_i, X_j | \Theta_{TT} = \theta)], \tag{5}$$

where Θ_{TT} is the test composite, X_i and X_j are scores on items i and j, and

$$\delta_{ij} = \begin{cases} 1 & \text{if items } i \text{ and } j \text{ are in the same cluster of } P \\ -1 & \text{otherwise.} \end{cases} \tag{6}$$

The index $D(P)$ is a measure of the degree of multidimensionality present in the partition P. It is obvious that numerous ways exist to partition items of a test into clusters, and each partition produces a value of $D(P)$. Let P^* be a partition such that $D(P^*) = \max\{D(P) | P \text{ is a partition}\}$. Then P^* is treated as the optimal simple dimensionality structure of the test, and $D(P^*)$ is treated as the maximum amount of multidimensionality present in the test data. For example, for a purely unidimensional test, the optimal dimensionality structure of the test is that all the items will be partitioned into one single cluster, and $D(P^*)$ for the test will be close to 0. It has been shown by Zhang and Stout (1999b) that when there is a true simple structure underlying test data, $D(P)$ will be maximized only for the correct partition.

To determine if the partition P^*, which produced the maximum DETECT index $D(P)$, is indeed the correct simple structure of the test, we can use the following ratio:

$$R(P^*) = \frac{D(P^*)}{\tilde{D}(P^*)}, \tag{7}$$

where

$$\tilde{D}(P^*) = \frac{2}{n(n-1)} \times \sum_{1 \le i \le j \le n} |E[\mathrm{Cov}(X_i, X_j | \Theta_{TT} = \theta)]|. \tag{8}$$

When there is an approximate simple structure underlying test data, then the ratio $R(P^*)$ is close to 1. The extent to which $R(P^*)$ differs from 1 is indicative of the degree to which the test structure deviates from the simple structure.

Because the true ability of an examinee is unobservable, $E[\mathrm{Cov}(X_i, X_j | \Theta_{TT} = \theta)]$ of equation (5) cannot be computed directly but must be estimated using observable data. There are two natural estimates of $E[\mathrm{Cov}(X_i, X_j | \Theta_{TT} = \theta)]$:

$$\widehat{\mathrm{Cov}}_{ij}(T) = \sum_{m=0}^{N} \frac{J_m}{J} \widehat{\mathrm{Cov}}(X_i, X_j | T = m), \tag{9}$$

where the conditional score $T = \sum_{l=1}^{N} X_l$ is the total score of all test items, J is the total number of examinees, and J_m is the number of examinees in subgroup m with the total score $T = m$. The other is the estimator based on the total score of remaining items given by

$$\widehat{\mathrm{Cov}}_{ij}(S) = \sum_{m=0}^{N-2} \frac{J_m}{J} \widehat{\mathrm{Cov}}(X_i, X_j | S = m), \tag{10}$$

where the score $S = \sum_{l=1, l \ne i, j}^{N} X_l$ is the total score of the remaining items, other than items i and j, and J_m is the number of examinees in subgroup m with the conditional score $S = m$.

When a test is unidimensional, $\widehat{\mathrm{Cov}}_{ij}(T)$ tends to be negative because items X_i and X_j are part of T. Therefore, $\widehat{\mathrm{Cov}}_{ij}(T)$ as an estimator of $E[\mathrm{Cov}(X_i, X_j | \Theta_T = \theta)]$ results in a negative bias (Junker, 1993; Zhang & Stout, 1999a). On the other hand, $\widehat{\mathrm{Cov}}_{ij}(S)$ tends to be positive and results in a positive bias (Holland & Rosenbaum, 1986; Rosenbaum, 1984; Zhang & Stout, 1999a).

Because $\widehat{\mathrm{Cov}}_{ij}(T)$ tends to have a negative bias and $\widehat{\mathrm{Cov}}_{ij}(S)$ tends to have a positive bias as estimators of $E[\mathrm{Cov}(X_i, X_j | \Theta_T = \theta)]$ in the unidimensional case, Zhang and Stout (1999b) proposed an average of these two estimates, resulting in the following index as an estimator of the theoretical DETECT index $D(P)$:

$$D_{ZS}(P) = \frac{2}{n(n-1)} \sum_{1 \le i \le j \le N} \delta_{ij} \widehat{\mathrm{Cov}}^*_{ij}, \tag{11}$$

where

$$\widehat{\mathrm{Cov}}^*_{ij} = \frac{1}{2} [\widehat{\mathrm{Cov}}_{ij}(S) + \widehat{\mathrm{Cov}}_{ij}(T)]. \tag{12}$$

An estimate of $R(P)$ can be similarly obtained. The DETECT software adopts a special technique, called the genetic algorithm, to divide items of a test into different dimensional clusters. The genetic algorithm iteratively mutates items to different dimensional clusters until the maximum degree of multidimensionality of the test $D_{\max}$, an estimate of $D(P^*)$, is obtained. The dimensional cluster pattern that produces $D_{\max}$ is treated as the final dimensionality structure of the test. The process is accelerated when the initial cluster solution for the genetic algorithm is obtained via cluster analysis developed by Roussos, Stout, and Marden (1993).

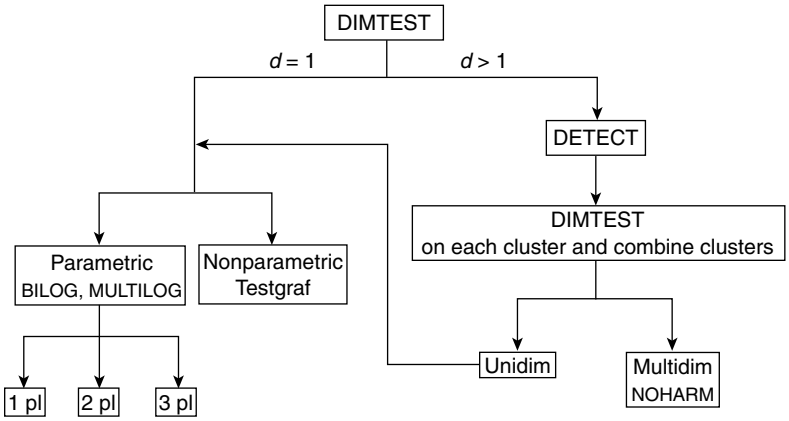
To interpret the results of DETECT in applications, Zhang and Stout (1999b) provided the following rule of thumb based on simulation studies. Divide the examinee sample into two parts: sample1 and sample2 (cross validation sample). Using sample1, find item partition, P_1^*, that maximizes the detect index for sample1, called $D_{\max}$. Using sample2, find P_2^*, that maximizes the detect index for sample2. Then using the item partition P_2^*, from the cross validation sample, compute the detect value for sample1, called D_{ref}. Generally is less than or equal to $D_{\max}$. A test is judged to be essentially unidimensional if D_{ref} is less than 0.1 or $\frac{D_{\max} - D_{ref}}{D_{ref}} > .5$.

5.4. Data Modeling

An algorithm is proposed below to model test data. As emphasized hitherto, the goal is to determine if unidimensional scoring is meaningful for given data. Although any appropriate methodology can be used to carry out the steps of the algorithm, DIMTEST and DETECT are recommended, as they are specifically developed for this purpose, easy to use, and nonparametric.

The flowchart in Figure 5.3 details the steps for test modeling, which are described in the algorithm following the flowchart. These steps are illustrated through the analyses of simulated data in the following section.

Figure 5.3 Flowchart Describing Steps for Test Modeling



5.4.1. An Algorithm for Test Modeling

Step 1. Use DIMTEST to determine if dimensionality, d, underlying test data is essentially 1.

Step 2. If $d = 1$, then fit a unidimensional model to data. Choose an appropriate unidimensional model. Exit.

Step 3. If $d > 1$, then investigate if test items can be decomposed into unidimensional clusters using DETECT.

Step 4. Test each cluster using DIMTEST to determine if $d = 1$.

Step 5. Combine clusters, if necessary, based on expert opinion and item content of the AT subtest of DIMTEST. Again test the hypothesis $d = 1$.

Step 6. If $d = 1$, go to Step 2. If $d > 1$ for any of the clusters, either delete them from the test or explore multidimensional modeling.

If unidimensional modeling is appropriate either on the whole test or on subtests (Step 2), one can fit either a parametric model or a nonparametric model. If a parametric model is desired, there are several models to choose from. Some of the commonly used models are the one-parameter logistic model (1PL), the two-parameter logistic model (2PL), or the three-parameter logistic model (3PL). Parameters of these models can be estimated using standard computer software such as BILOG (Mislevy & Bock, 1989), MULTILOG (Thissen, 1991), and RUMM (Sheridan, Andrich, & Luo, 1998). For more detailed information about fitting various parametric models, estimating parameters, and scoring, refer to Embretson and Reise (2000) and Thissen and Wainer (2001). An alternative is nonparametric modeling. Nonparametric estimation of item response functions can be carried out using the software TESTGRAF (Douglas & Cohen, 2001; Ramsay, 1993). If unidimensional modeling is not appropriate either for the whole test or after splitting into subtests (Step 6), multidimensional modeling of data is necessary. Currently, multidimensional models and estimation of their parameters are limited. One program that has shown a lot of promise in estimating multidimensional parameters is NOHARM (Fraser, 1986). For details about fitting multidimensional models, see Reckase (1997), McDonald (1997), and Ackerman, Neustel, and Humbo (2002).

5.4.2. Illustration of Test Modeling

Data modeling will be illustrated using simulated data. Unidimensional and two-dimensional data were simulated. All data sets had 30 items and 2,000 examinees, which are typical values usually encountered in applications. One unidimensional test and four two-dimensional tests were generated. Unidimensional data were generated using a unidimensional two-parameter logistic model (Hambleton & Swaminathan, 1985).

$$P_i(\theta_j) = \frac{1}{1 + \exp[-1.7[a_i(\theta_j - b_i)]]}, \qquad (13)$$

where $P_i(\theta_j)$ is the probability of a correct response to the dichotomous item i by an examinee with ability (θ_j), a_i is the discrimination parameter of the dichotomous item i, and b_i is the difficulty parameter of item i.

Figure 5.4 Item Vectors Representing the Simple Structure Test

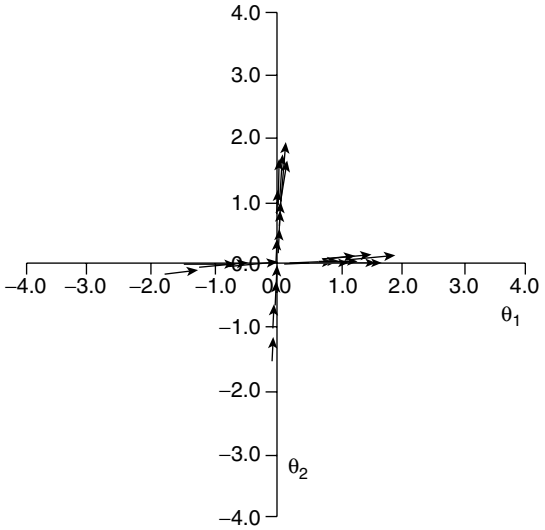
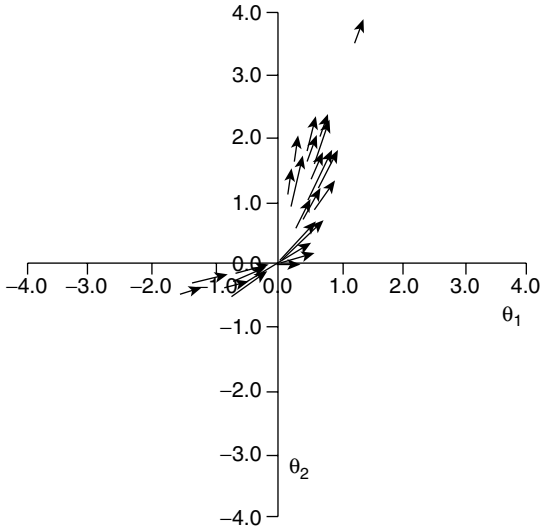


Figure 5.5 Item Vectors Representing the Complex Structure Test



Examinee abilities were randomly generated from the standard normal distribution with mean 0 and the standard deviation 1. Item parameters were randomly selected from a pool of parameter estimates from several nationally administered standardized achievement tests.

Two types of two-dimensional data were generated: simple structure and complex structure. Item parameters for the simple structure were such that items of each dimension were located within 15 degrees from the respective axes, as illustrated in Figure 5.4. Item parameters for the complex structure were selected from a two-dimensional calibration of an American College Test (ACT) mathematics test in which items span the entire two-dimensional space, as illustrated in Figure 5.5.

Two levels of correlation between dimensions (ρ_{θ_1,θ_2}) were considered: .5 and .7. This resulted in four two-dimensional tests: simple structure with $\rho = .5$, simple structure with $\rho = .7$, complex structure with $\rho = .5$, and complex structure with $\rho = .7$. For each two-dimensional test, the first half of the items (Items 1 to 15) measured predominantly the first dimension, and the second half measured predominantly the second dimension. Each examinee's abilities θ_1 and θ_2 were randomly generated from a bivariate normal distribution with an appropriate correlation coefficient between the abilities. Two-dimensional data were generated using the following two-dimensional, two-parameter compensatory model (Reckase, 1997; Reckase & McKinley, 1983):

$$P_i(\theta_{1j}, \theta_{2j}) = \frac{1}{1 + \exp[-1.7(a_{1i}\theta_{1j} + a_{2i}\theta_{2j} + b_i)]}, \quad (14)$$

where $P_i(\theta_{1j}, \theta_{2j})$ is the probability of a correct response to the dichotomous item i by an examinee j with ability $(\theta_{1j}, \theta_{2j})$, a_{1i} is the discrimination parameter of the dichotomous item i on dimension θ_1, a_{2i} is the discrimination parameter of the item i on dimension θ_2, and b_i is the difficulty parameter of item i. The simulated data sets are described in Table 5.1.

5.4.3. Results of Data Analyses

For each data set, the correct model was arrived at by following the steps described in the algorithm for test modeling, as illustrated in Figure 5.3. Results of the analyses are tabulated in Tables 5.2 and 5.3. These results will be summarized below in detail for each of the tests.

Uni.dat: DIMTEST results ($T = 0.85$ and $p = .20$) showed that it is essentially unidimensional. Hence, unidimensional modeling is appropriate for these data.

Table 5.1 Description of Simulated Data

Test	*# Items*	*# Examinees*	ρ^a	*Dimensionality*
uni.dat	30	2,000	—	$d = 1$
simplr5.dat	30	2,000	0.5	$d = 2$, simple structure
simplr7.dat	30	2,000	0.7	$d = 2$, simple structure
realr5.dat	30	2,000	0.5	$d = 2$, complex structure
realr7.dat	30	2,000	0.7	$d = 2$, complex structure

a. Denotes the correlation between latent abilities for two-dimensional tests.

Table 5.2 DIMTEST and DETECT Results

	DIMTEST		*DETECT*			
Test	*T*	*p*	D_{max}	*R*	*# Clusters*	*Item Clusters*
uni.dat	0.85	.20	—	—	—	—
simplr5.dat	9.69	.00	1.33	0.98	2	1–15, 16–30
simplr7.dat	6.0	.00	1.58	0.74	2	1–15, 16–30
realr5.dat	2.63	.00	0.16	0.29	3	(1, 4, 6, 7, 10, 11, 13, 14, 15, 27); (2, 5, 8, 9, 12, 19, 23, 29); (3, 16, 17, 18, 20, 21, 22, 24, 25, 26, 28, 30)
realr7.dat	0.86	.19	—	—	—	—

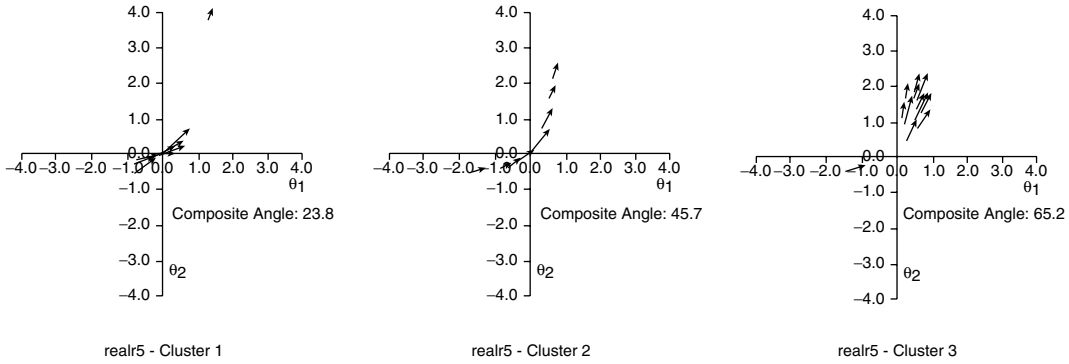
Table 5.3 Further Analyses of Two-Dimensional Data

		DIMTEST		*DETECT*	
Test	*Item Cluster*	*T*	*P*	D_{max}	*R*
simplr5.dat	1–15	–0.77	.78	—	—
	16–30	0.03	.49	—	—
simplr7.dat	1–15	–1.36	.91	—	—
	16–30	0.90	.18	—	—
realr5.dat	1, 4, 6, 7, 10, 11, 13, 14, 15, 27	–.76	.78	—	—
	2, 5, 8, 9, 12, 19, 23, 29	—	—	0.01	0.02
	3, 16, 17, 18, 20, 21, 22, 24, 25, 26, 28, 30	1.04	.15		
realr5.dat clusters 1 and 2	1, 2, 4 to 15, 19, 23, 27, 29	0.52	.30	—	—

simplr5.dat: DIMTEST results ($T = 9.69$ and $p = .00$) indicated the presence of more than one dominant dimension underlying these test data. DETECT analyses resulted in a two-cluster solution with a high value of $D_{\max}$ (1.33) and an R-value close to 1, indicating two dimensions with a simple structure solution. As expected, Items 1 to 15 formed one cluster, and the rest of the items formed the second cluster. Further analyses on these clusters, shown in Table 5.3, showed that each of these clusters is unidimensional ($T = -0.77$ and $p = .78$ for Items 1 to 15; $T = 0.03$ and $p = .49$ for Items 16 to 30). Hence, these subtests are amenable to unidimensional modeling.

simplr7.dat: These test data were also assessed as multidimensional ($T = 6.0$ and $p = .00$) by DIMTEST. DETECT analyses on these data resulted in a two-cluster solution. However, the $D_{\max}$ (0.58) and R-values (0.74) were not high, indicating that the simple structure solution is not as explicit as it was for simplr5.dat. This is due to high correlation between latent abilities. Nonetheless, it is noteworthy that DETECT was able to correctly classify items into clusters given the high degree of correlation between abilities. Further analysis on the clusters, shown in Table 5.3, showed that each cluster is unidimensional ($T = -1.36$ and $p = .91$ for Items 1 to 15; $T = 0.90$ and $p = .18$ for Items 16 to 30).

Figure 5.6 Item Vectors Representing the Three Clusters in the Test: realr5



realr5.dat: DIMTEST results ($T = 2.63$ and $p = .00$) indicated that the data violated the unidimensionality assumption. Subsequent DETECT analyses showed three clusters. Although the DETECT procedure split the test items into three clusters, the corresponding $D_{\max}$ (0.16) and R-values (0.29) were small, indicating that the degree of multidimensionality was not of a concern. In fact, the $D_{\max}$ value was within the range of what is expected for a unidimensional test. Here, the unidimensionality assumption is violated. However, there is not enough evidence of multidimensionality to warrant significant separate clusters.

To understand the nature of multidimensionality, each of the clusters was further analyzed for unidimensionality using DIMTEST. As the results suggest in Table 5.3, Clusters 1 and 3 were confirmed as unidimensional by DIMTEST ($T = -0.76$ and $p = .78$ for Cluster 1; $T = 1.04$ and $p = .15$ for Cluster 3). Because Cluster 2 contained too few items to apply DIMTEST, its dimensionality was estimated using DETECT. Note that the $D_{\max}$ value (0.01) associated with Cluster 3 was very small and resembles a value associated with unidimensional tests. Hence, one may treat this cluster as unidimensional.

DIMTEST also provided clues regarding the source of the multidimensionality. If the null hypothesis of $d = 1$ is rejected, it means that items in the subtest AT are contributing to multidimensionality. Upon observing the AT subtest of DIMTEST results of realr5.dat, it was found that there was an overlap of items between Cluster 3 and the AT subtest. Hence, it was conjectured that Cluster 3 was dimensionally distinct from Clusters 1 and 2. Hence, Clusters 1 and 2 were combined to confirm if the combined subtest is unidimensional. DIMTEST analysis confirmed unidimensionality of this subtest ($T = 0.52$ and $p = .30$). Hence, there are two unidimensional subtests of realr5.dat.

Figure 5.6 shows a graphical display of vector plots of items in the three clusters identified by DETECT. Contrasting Figures 5.5 and 5.6, it can be seen that the item vectors in Figure 5.5 (in which abilities have a correlation of 0.5) are split into three clusters by the DETECT procedure. The test composite vector of Cluster 1 is at 23.8 degrees from the θ_1-axis, the test composite vector of Cluster 2 is at 45.7 degrees from the θ_1-axis, and the test composite vector of Cluster 3 is at 65.2 degrees from the θ_1-axis. Both the DIMTEST and DETECT procedures are sensitive to the differences among these three clusters. As the detailed analyses revealed, Clusters 1 and 2 can be combined to form a unidimensional subtest, whereas Cluster 3 is an independent cluster dimensionally different from the other two clusters.

realr7.dat: DIMTEST analyses of this test revealed unidimensionality ($T = 0.86$ and $p = .19$). This is not surprising as the items span the entire two-dimensional space in which the two abilities are highly correlated. Hence, this group of items is best captured by a unidimensional vector encompassing all items in the space. Unidimensional scoring is the best way to summarize these data.

In summary, unidimensional modeling was appropriate for the following test data: uni.dat and realr7.dat. The former is an inherently unidimensional test, whereas the latter resembles a unidimensional test because of high correlation between abilities coupled with items spanning the entire two-dimensional space, as in Figure 5.5. For both of these tests, the DIMTEST results indicated unidimensionality. Two-dimensional

data sets—simplr5.dat and simplr7.dat, both simple-structure tests—were assessed as multidimensional based on DIMTEST analyses. DETECT results confirmed this fact by indicating a high degree of multidimensionality, as evidenced by large $D_{\max}$ and R-values. It is remarkable that DETECT, despite highly correlated abilities for simplr7.dat, correctly partitioned test items into clusters/subtests. The subtests of simplr5.dat and simplr7.dat were further assessed by DIMTEST as unidimensional. Hence, unidimensional modeling for each of these subtests is meaningful. Among all simulated test data, the dimensionality structure of realr5.dat turned out to be the most complex. For these test data, even though the DIMTEST analyses indicated the presence of multidimensionality, DETECT analyses indicated a very low degree of multidimensionalily. Further investigation and detection of the source of multidimensionality in realr5.dat led to the identification of two subtests, which were each unidimensional. Hence, all three two-dimensional tests could be split into subtests for unidimensional modeling or could be combined for two-dimensional modeling and scoring.

5.5. Summary and Conclusions

The aim of a test is to accurately capture the examinee's position on a continuum of latent trait(s) of interest. To accomplish this, one must use a model that best explains given data, which is an interaction between items and the examinee population taking the test. Most commonly used models to explain test data comprise monotone, local independent, and unidimensional assumptions. However, increasingly, tests are designed to measure more than one dominant trait. Hence, it has become ever more important to empirically investigate the suitability of the unidimensional modeling of test data. This chapter has provided a modeling algorithm using a series of procedures to investigate whether test data are amenable to monotone, local independent, and unidimensional modeling. The proposed algorithm for test modeling was illustrated using simulated test data. Although the algorithm described here provides a framework for test modeling, the process is more of an art than a science. Often, data in the real world may not strictly satisfy the criteria proposed here for test modeling. For example, results of DIMTEST and DETECT may lead to conclusions that test data do not adhere to unidimensional modeling. At the same time, test data may not warrant multidimensional modeling (e.g., realr5.dat). In such a situation, it is important to go beyond statistical analyses and consult content experts and test specifications to decide the most appropriate modeling of test data. Clearly, modeling test data involves many decisions and thus is more a craft than an exact science.

Another important aspect of test modeling is to consider implications of dimensionality considerations. There are well-established methodologies and a choice of software for fitting unidimensional models and estimating parameters of items and examinees. Hence, the selection of multidimensional models over unidimensional models needs careful examination. Other important factors to consider are the cost, improvement in accuracy and understanding of the results, and communication of results with the public.

References

Ackerman, T. (1996). Graphical representation of multidimensional item response theory. *Applied Psychological Measurement, 20,* 311–329.

Ackerman, T. A., Neustel, S., & Humbo, C. (2002, April). *Evaluating indices used to access the goodness-of-fit of the compensatory multidimensional item response model.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Douglas, J. A. (1997). Joint consistency of nonparametric item characteristic curve and ability estimation. *Psychometrika, 62,* 7–28.

Douglas, J. A., & Cohen, A. (2001). Nonparametric ICC estimation to assess fit of parametric models. *Applied Psychological Measurement, 25,* 234–243.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum.

Fraser, C. (1986). *NOHARM: An IBM PC computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory.* Armidale, Australia: University of New England.

Froelich, A. G. (2000). *Assessing unidimensionality of test items and some asymptotics of parametric item response theory.* Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Amsterdam: Kluwer Nijhoff.

Hattie, J., Krakowski, K., Rogers, J., & Swaminathan, H. (1996). An assessment of Stout's index of essential dimensionality. *Applied Psychological Measurement, 20,* 1–14.

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

Humphreys, L. G. (1985). General intelligence: An integration of factor, test, and simplex theory. In B. B. Wolman (Ed.), *Handbook of intelligence* (pp. 201–224). New York: John Wiley.

Humphreys, L. G. (1986). An analysis and evaluation of test and item bias in the prediction context. *Journal of Applied Psychology, 71,* 327–333.

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, University of Illinois, Urbana-Champaign.

McDonald, R. P. (1997). Normal-ogive multidimensional model. In W. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 258–269). New York: Springer-Verlag.

Mislevy, R. J., & Bock, R. D. (1989). *BILOG: Item analysis and test scoring with binary logistic models.* Chicago: Scientific Software.

Nandakumar, R. (1993). Assessing essential unidimensionality of real data. *Applied Psychological Measurement, 1,* 29–38.

Nandakumar, R. (1994). Assessing latent trait unidimensionality of a set of items: Comparison of different approaches. *Journal of Educational Measurement, 31,* 1–18.

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

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56,* 611–630.

Ramsay, J. O. (1993). *TESTGRAF: A program for the graphical analysis of multiple choice test and questionnaire data: TESTGRAF user's guide.* Montreal, Quebec: Department of Psychology, McGill University.

Reckase, M. D. (1997). A liner logistic multidimensional model for dichotomous item response data. In W. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271–286). New York: Springer-Verlag.

Reckase, M. D., & McKinley, R. L. (1983, April). *The definition of difficulty and discrimination for multidimensional item response theory models.* Paper presented at the annual meeting of American Educational Research Association, Montreal, Quebec.

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

Roussos, L. A., Stout, W. F., & Marden, J. I. (1993, April). *Dimensional and structural analysis of standardized tests using DIMTEST with hierarchical cluster analysis.* Paper presented at the annual NCME meeting, Atlanta, GA.

Seraphine, A. E. (2000). The performance of the Stout T procedure when latent ability and item difficulty distributions differ. *Applied Psychological Measurement, 24,* 82–94.

Sheridan, B., Andrich, D., & Luo, G. (1998). *RUMM: Rasch unidimensional measurement models.* Duncraig, Australia: RUMM Laboratory.

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

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55,* 293–325.

Stout, W. (2002). Psychometrics: From practice to theory and back. *Psychometrika, 67,* 485–518.

Stout, W., Froelich, A. G., & Gao, F. (2001). Using resampling to produce an improved DIMTEST procedure. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 357–376). New York: Springer-Verlag.

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

Thissen, D. (1991). *MULTILOG user's guide* – (Version 6). Chicago: Scientific Software.

Thissen, D., & Wainer, H. (Eds.). (2001). *Test scoring.* Mahwah, NJ: Lawrence Erlbaum.

Tucker, L. R., Koopman, R. F., & Linn, R. L. (1969). Evaluation of factor analytic research procedures by means of simulated correlation matrices. *Psychometrika, 34,* 421–459.

van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory.* New York: Springer-Verlag.

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., & 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.