LOCAL ITEM DEPENDENCE FOR ITEMS ACROSS TESTS CONNECTED BY COMMON STIMULI

WEN-CHUNG WANG
National Chung Cheng University

YING-YAO CHENG
National Sun Yat-sen University

MARK WILSON
University of California at Berkeley

A parallel design, in which items across different scales within an instrument share common stimuli and subjects respond to the common stimulus for each scale, is sometimes used in questionnaires or inventories. Because the items across scales share the same stimuli, the assumption of local item independence may not hold, thereby violating the assumption of local item independence under standard psychometric models. In this study, the authors describe a multidimensional item response model to detect specific forms of local item dependence. Three real data sets were analyzed to illustrate implications and applications of the proposed method.

Keywords: *test design; item bundle; testlet; multidimensional item response model; Rasch model; item response theory*

Item response theory (IRT) has been widely used to model data of educational and psychological tests. A basic postulate of IRT is that the performance of an examinee on an item can be explained by a function of a set of latent variables associated with persons and a set of item parameters (e.g., item difficulty and item discrimination). Whenever several items are administered simultaneously, the assumption of local item independence is made for effec-

Correspondence concerning this article should be addressed to Wen-Chung Wang, Department of Psychology, National Chung Cheng University, Chia-Yi, Taiwan; e-mail: psywcw@ccu.edu.tw.

Educational and Psychological Measurement, Vol. 65 No. 1, February 2005 5-27
DOI: 10.1177/0013164404268676

tive theoretical work, such as forming likelihood equations to derive estimators. Local item independence means that for any group of examinees all characterized by the same values of the latent variable(s), the conditional distribution of the item characteristic curves are all independent of one another. Implicitly, it states that the latent variables under consideration span the complete latent space. Local item independence does not suggest that item scores are unrelated to one another for the whole group of examinees. What it really means is that item scores are related to one another only through the latent variables. If the values of an examinee's latent variables are already known, any knowledge of his or her performances on other items will add nothing to this determination. If it does add something, then performance on the items would depend in part on some variables other than the target variables, which is contrary to the assumption of local item independence (Lazarsfeld & Henry, 1968; Lord, 1980; Lord & Novick, 1968).

If the assumption of local item independence is violated, any statistical analysis based on it would be misleading. Specifically, estimates of the latent variables and item parameters will generally be biased because of model misspecification, which in turn leads to incorrect decisions on subsequent statistical analysis, such as testing group differences and correlations between latent variables. In addition, it is not clear what constructs the item responses reflect, and consequently, it is not clear how to combine those responses into a single test score, whether IRT is being used or not (Steinberg & Thissen, 1996).

There are a number of situations in which the assumption of local item independence may be questionable. First, consider filling out an inventory on which the first item might ask to what extent one has a certain characteristic in general, such as overall appearance (or dominance, self-esteem, anxiety, etc.). The following items might ask to what extent one has that certain characteristic in some specific situations, such as face, chest, arms, and legs, with respect to appearance, or contexts such as home or work for dominance, and so on. It seems likely that an examinee would tend to respond in a consistent way across the several questions, and in particular, the responses may well be more consistent than if the questions had been asked in a less time- and place-dependent way. Second, consider an examinee taking ability test items with the same format, which is rather unfamiliar to the examinee. The examinee may find the items relatively more difficult, especially the first few items, until he or she becomes more familiar with the novel format. The responses to the later items may be affected by those to the previous ones, because of a carryover effect as the examinee learns about the format while taking the test.

Third, consider a problem-solving question that calls for multiple steps (or subitems), so that knowing the answer to one item increases the chances of knowing the answer to the next one. In contrast, an error in one item may cause further errors in the subsequent items. Another testing situation similar to this is the cloze procedure in language tests, whereby examinees are asked

to fill in the blanks of missing words in a paragraph. An error in filling in a missing word in one place might cause further errors in others. This item-chaining effect is a threat to the assumption of local item independence. Fourth, consider that items in a test are clustered and connected by attributes such as common stimuli, item stems, item structures, or item contents. Several terms have been used to denote such item subsets, such as *testlets* (Wainer & Kiely, 1987) or *item bundles* (Rosenbaum, 1988). Item bundles have been used in many in educational psychological tests, such as language reading comprehension, performance-based items in science and mathematics, and scenario-based measures. Local item dependence can be produced by an examinee's unusual level of interest or background knowledge about the common stimuli or by the fact that information used to answer different items is interrelated in the stimuli.

Issues of local item dependence within the IRT framework have been well documented in the literature (Andrich, 1985; Bradlow, Wainer, & Wang, 1999; Chen & Thissen, 1997; Hoskens & De Boeck, 1997; Jannarone, 1986, 1995; Rosenbaum, 1984, 1988; Sireci, Thissen, & Wainer, 1991; Thissen, Steinberg, & Mooney, 1989; Wainer, 1995; Wainer & Lukhele, 1997; Wainer, Sireci, & Thissen, 1991; Wainer & Thissen, 1996; Wainer & Wang, 2000; Wang, Bradlow, & Wainer, 2002; Wilson, 1988; Wilson & Adams, 1995; Yen, 1984, 1993). Fitting standard item response models to sets of interdependent items (item bundles) tends to overestimate the precision of person measures and yield biased estimation for item parameters. The overstatement of precision and biased estimation lead to inaccurate inferences about the parameters. A method for analyzing an item bundle is to treat it as a single superitem, score it polytomously, and apply polytomous item response models such as the partial credit model (Masters, 1982), the graded response models (Samejima, 1969), or the generalized partial credit model (Muraki, 1992). This approach might be appropriate when the local dependence between items within a bundle is moderate and the test contains a large proportion of independent items (Wainer, 1995). Using this approach, as long as the total scores in the bundle are identical, they will be assigned to the same category. In this case, the information in the exact response patterns within the bundle cannot be retrieved.

Another method of accounting for local item dependence is to consider the response patterns of an item bundle as a basic unit instead of responses to single items. For example, Rosenbaum (1984, 1988) introduced the idea of bundle independence as a way to address the issues of local dependence among items within bundles. He proposed that instead of expressing the likelihood of a response vector as a product of the probabilities of responses to individual items, the likelihood is expressed as a product of the probabilities of responses to the bundles within which one might expect local dependence. Local bundle independence, rather than local item independence, is assumed in the model. To test whether items within a bundle show local item depend-

ence, one could use the likelihood ratio test to compare an augmented bundle model, in which items within a bundle are assumed to be locally dependent, with a reduced bundle model, in which they are assumed to be locally independent.

The research works listed above focus mainly on using unidimensional item response models to detect local item dependence, because the dependence is suspected to occur within tests. Using the same bundle (item pattern) methodology, Hoskens and De Boeck (2001) presented multidimensional item response models for complex cognitive tasks for which more than one latent variable is involved. In this study, we consider a particular testing situation, in which unidimensional item response models are not applicable, because common stimuli are shared by items in different tests that are designed to measure different latent variables. This calls for multidimensional item response models, because more than one latent variable is involved. To further clarify the meaning of local item dependence that arises when items across tests share common stimuli, take the following example. Suppose one is asked to measure people's heights and weights with rulers and scales. Because rulers and scales do not interfere with each other, the measurements of the two variables, height and weight, are independent. (This measurement independence should not be confused with the high correlation between height and weight.) Suppose then that one is asked to judge people's heights and weights without rulers and scales but with his or her eyes. Will the judgments of height and weight be independent? Unfortunately, the judgments would be mutually dependent because one might overestimate (or underestimate) people's weights because they appear to be tall (or short). This dependence, being analogous to the local item dependence that arises when items across tests are connected by common stimuli, will contaminate the measurement and the relationship between height and weight that one might be interested in most.

In the following, we describe a particular test design strategy that is sometimes used in personality inventories or questionnaires and may cause local dependence among items relating to subscales within the same instrument. Several nested multidimensional item response models for depicting local item dependence are proposed. Likelihood ratio tests are used to test whether local item dependence occurs and whether the patterns of local item dependence are identical across common stimuli. An empirical example with three data sets is given to illustrate applications and implications of the proposed method.

Example: Design of the Creativity Development Inventory

Consider the following example, in which the particular test design contains items across different tests connected by common stimuli, thus causing the potential for local item dependence. Cheng and Wang (1999) designed

The Parallel Design

Judge the following personal characteristics in terms of (a) how important are they to creativity development, and (b) how much you possess.

	Importance			Possession		
	very little	not much	very much	very little	not much	very much
1. Multiple positive personalities	☐	☐	☐	☐	☐	☐
2. High self-motivation	☐	☐	☐	☐	☐	☐
⋮						
14. Tolerance to ambiguous conditions	☐	☐	☐	☐	☐	☐
15. Strong confidence	☐	☐	☐	☐	☐	☐

The Sequential Design

Judge the following personal characteristics in terms of how important to creativity development.

	Importance		
	very little	not much	very much
1. Multiple positive personalities	☐	☐	☐
⋮			
15. Strong confidence	☐	☐	☐

Judge the following personal characteristics in terms of how much you possess.

	Possession		
	very little	not much	very much
1. Multiple positive personalities	☐	☐	☐
⋮			
15. Strong confidence	☐	☐	☐

Figure 1. The Creativity Development Inventory with the parallel and sequential designs.

the Creativity Development Inventory, which contains 15 personal characteristics. Subjects were asked to judge how important the characteristics are to creativity development and how much they possess the characteristics, both on a 3-point scale: *very little*, *not much*, and *very much*. For the sake of subjects' response convenience and economy of space, these characteristics were printed on the left, and the two scales, perceived importance and perceived possession, were printed in parallel on the right, as shown in the upper part of Figure 1. This design, called the parallel design, made the subjects

judge the first characteristic in terms of importance, followed immediately by the degree of possession, then move on to the second characteristic, and so on. A standard way to analyze these kinds of data is simply to report the raw sum of each scale as the subject's score and the Pearson correlation between the two sums as the relationship between perceived importance and perceived possession. In doing so, one assumes implicitly that items in each scale measure solely a single latent variable and that these items are locally independent both within and between scales.

This implicit assumption, being analogous to the local item independence, may not hold in this particular example. When a subject judges possession, his or her response to importance is so recent that it may be affected by the previous responses. Carryover effects are thus expected; that is, the subject might overestimate (or underestimate) the degree of possession when he or she considers the characteristic important. If the assumption of local item independence does not hold, as previously described, the meanings of the two latent variables and the derived measures will be contaminated, and the greater the dependence, the more severe the contamination will be.

The parallel design is occasionally used in questionnaires or inventories. For instance, children are asked how they perceive their fathers and mothers separately (e.g., "Is Father warm, kind, outgoing, patient, etc.?" and "Is Mother warm, kind, outgoing, patient, etc.?"). Students may be requested to describe how the statements match their learning strategies when they are working on (a) languages, (b) mathematics, and (c) sciences, separately. Clients may be asked to rate how the statements describe their anger reactions (a) within this month, (b) within this week, and (c) today, separately. Clients may be asked to rate various life stress events on multiple aspects, such as frequency and intensity. Participants may be asked to judge how much they own self-esteem–related characteristics (e.g., health, wealth, reputation, family support, friendship, etc.) and how important these characteristics are to their overall self-esteem. Because the statements about the targets (e.g., father and mother) are identical, the parallel design is used for the convenience of response and economy of space. Scenario-based measures, in which each scenario is assessed with multiple aspects, are further examples of the parallel design.

The parallel design can be viewed in some sense as a kind of within-subjects designs in which differential transfer would be even more serious because subjects have to switch their attention back and forth across aspects (e.g., perceived importance and perceived possession in this example). One way of reducing the dependence is to adopt a sequential design, in which the characteristics together with the first scale as a whole are printed and then followed by the same characteristics together with the second scale as a whole, as illustrated in the lower part of Figure 1. With this sequential design, subjects are forced to respond to all the characteristics in terms of importance

and respond to them again in terms of possession. Subjects do not have to switch their attention back and forth. Even better, subjects could be asked to conduct some distracting tasks between responding to the two scales to further reduce the dependence. If the delay is long or the distracting tasks function well, the dependence would be even more attenuated. However, it is not safe to ignore local item dependence in the sequential design without checking.

Method

Unidimensional item bundle models are not appropriate for detecting local item dependence in the parallel design, because the dependence is suspected to exist between scales rather than within scales. However, the idea of local bundle independence still applies. One could form a bundle for items that are connected by the same stimulus across tests. In the creativity example, two 3-point scales are used, so each bundle has nine possible response patterns, one for each score pattern. Instead of modeling the probabilities of item responses in a test, we analyze the bundles. The goals are to detect whether local item dependence exists and whether the patterns of dependence across bundles are identical.

To achieve these goals, three nested multidimensional bundle models are formed: (a) the independence bundle model, in which items within each scale are assumed to measure solely a single latent variable and are locally independent; (b) the saturated-dependence bundle model, in which items in different scales that share the common stimulus may interfere with one another; and (c) the uniform-dependence bundle model, in which the patterns of dependence are constrained to be identical across bundles. Because the three models are nested (to be shown later), likelihood ratio tests can be used to compare the difference between the saturated-dependence bundle model and the independence bundle model to check whether local item dependence occurs and to compare the difference between the saturated-dependence bundle model and the uniform-dependence bundle model to check whether the patterns of dependence are identical across bundles.

Take the item responses in the Creativity Development Inventory (see Figure 1) as an example. In the independence bundle model, the joint probability of score pattern in bundle b is modeled as

$$\log\left(\frac{p_b(l,m)}{p_b(l-1,m)}\right) = \theta_1 - \delta_{1bj}, \tag{1}$$

$$\log\left(\frac{p_b(l,m)}{p_b(1,m-1)}\right) = \theta_2 - \delta_{2bj}, \tag{2}$$

where $p_b(l, m)$ denotes the joint probability of scoring l on Latent Variable 1 and scoring m on Latent Variable 2 for bundle b; θ_1 is and θ_2 are the levels of Latent Variables 1 and 2, respectively; and δ_{1bj} and δ_{2bj} are the jth-step difficulty of bundle b on Latent Variables 1 and 2, respectively. In this model, four item parameters are estimated for a bundle. These four parameters are in fact equivalent to those step parameters for the two corresponding items in the two scales when the partial credit model (Masters, 1982) is applied separately, one scale at a time. It can be shown that the joint probability of any score pattern under the independence bundle model is equal to the product of the corresponding two individual probabilities under the partial credit model.

For the family of Rasch models, the number of free item parameters in an item is at most the number of item categories minus one, because one category has to be the reference category for identification of the parameters. Given a bundle with nine response categories, at most eight item parameters could be estimated. In the saturated-dependence bundle model, all the eight item parameters are used to depict the local item dependence. The joint probability of score pattern in bundle b is modeled as

$$\log\left(\frac{p_b(l,m)}{p_b(l-1,m)}\right) = \theta_1 - \delta_{1bj} - \tau_{bk}, \tag{3}$$

$$\log\left(\frac{p_b(l,m)}{p_b(l,m-1)}\right) = \theta_2 - \delta_{2bj} - \tau_{bk}, \tag{4}$$

where τ_{bk} is added to depict the local item dependence and is thus referred to as the dependency parameter. In addition to the four step parameters, four dependency parameters (τ_{bk}, $k = 1, 2, 3, 4$), are added to model the dependence among the nine response categories within each bundle.

There are several models that are medial in complexity between the saturated-dependence bundle model and the independence bundle model. An interesting one is the model in which the patterns of dependence are identical across bundles, which is called the uniform-dependence bundle model. Specifically, τ_{b1}, τ_{b2}, τ_{b3}, and τ_{b4} are constrained to be identical across all 15 bundles:

$$\log\left(\frac{p_b(l,m)}{p_b(l-1,m)}\right) = \theta_1 - \delta_{1bj} - \tau_k, \tag{5}$$

$$\log\left(\frac{p_b(l,m)}{p_b(l,m-1)}\right) = \theta_2 - \delta_{2bj} - \tau_k, \tag{6}$$

where τ_{bk} in the saturated-dependence bundle model is now constrained to be τ_k. We suspect the patterns to be identical across all bundles, because Likert-type scales were used so that the subjects were using the same response labels across characteristics.

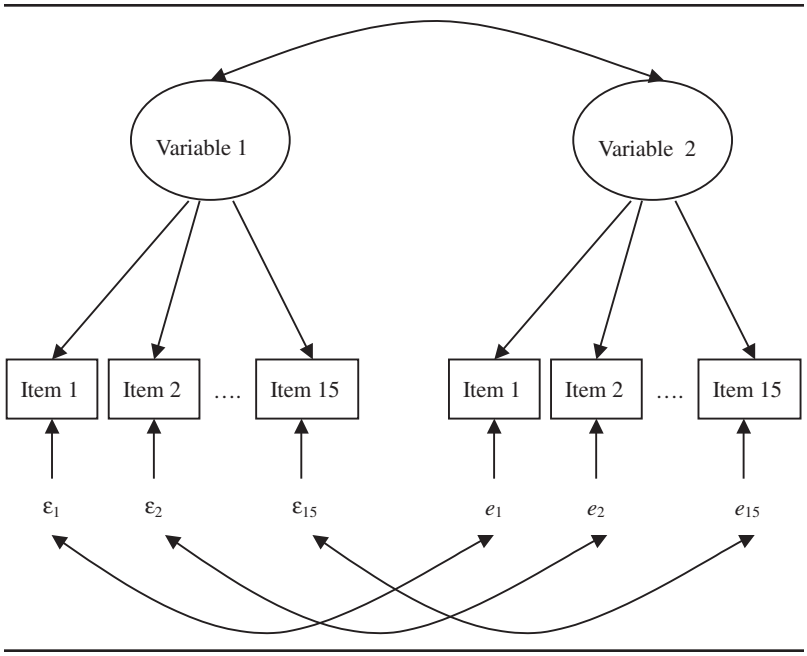

Figure 2. A graphical representation of local item dependence for items between two tests that are connected by common stimuli.

From the formulations for the three bundle models, one finds that the saturated-dependence bundle model (equations 3 and 4) contains the uniform-dependence bundle model (equations 5 and 6) as a special case by constraining the dependency parameters to be identical across bundles, which in turn contains the independence bundle model (equations 1 and 2) as a special case by constraining the dependency parameter to be zero. Figure 2 depicts intuitively a graphic representation of the local item dependence for items between two scales connected by common stimuli when the saturated-dependence bundle model is adopted. The error terms are connected pairwise to depict the local item dependence (e.g., the error term of Item 1 in Latent Variable 1 is connected to that in Latent Variable 2). In the uniform-dependence bundle model, all the 15 curves connecting two error terms are constrained to be identical. In the independence bundle model, the error terms are independent, and thus the curves connecting the error terms no longer exist.

The three multidimensional bundle models could not be implemented within the framework of unidimensional item response models, because more than one latent variable is involved in the probability equations (see equations 1 through 6). One could, however, apply the multidimensional random coefficients multinomial logit model (MRCMLM; Adams, Wilson, &

Wang, 1997) to calibrate the parameters in the three bundle models. Under the MRCMLM, which is a multidimensional extension of the random coefficients multinomial logit model (Adams & Wilson, 1996), person n's levels on the D latent variables are denoted as $\boldsymbol{\theta}_n' = (\theta_{n1}, \ldots, \theta_{nD})$, which are considered to represent a random sample from a population with a multivariate density function $g(\boldsymbol{\theta}, \boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ indicates a vector of parameters that characterize the distribution. If g is constrained to be normal, then $\boldsymbol{\alpha} \equiv (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The probability of a response in category j of item i for person n is

$$p_{nij} = \frac{\exp(\mathbf{b}_{ij}'\boldsymbol{\theta}_n + \mathbf{a}_{ij}'\boldsymbol{\xi})}{\sum_{u=1}^{k_i} \exp(\mathbf{b}_{iu}'\boldsymbol{\theta}_n + \mathbf{a}_{iu}'\boldsymbol{\xi})} \tag{7}$$

where k_i is the number of categories in item i, $\boldsymbol{\xi}$ is a vector of difficulty parameters that describe the items, $\mathbf{b}_{ij}$ is a score vector given to category j of item i across the D latent variables, and $\mathbf{a}_{ij}$ is a design vector given to category j of item i that describes the linear relationship among the elements of $\boldsymbol{\xi}$. Equation 7 can be expressed as

$$\begin{aligned} \log\left(\frac{p_{nij}}{p_{ni(j-1)}}\right) &= (\mathbf{b}_{ij}' - \mathbf{b}_{i(j-1)}')\boldsymbol{\theta}_n + (\mathbf{a}_{ij}' - \mathbf{a}_{i(j-1)}')\boldsymbol{\xi} \\ &\equiv \mathbf{b}_{ij}^{*\prime}\boldsymbol{\theta}_n + \mathbf{a}_{ij}^{*\prime}\boldsymbol{\xi} \end{aligned} \tag{8}$$

which is more consistent with the standard expression of the family of Rasch models and the above three bundle models. Notice that $\mathbf{a}_{ij}$ and $\mathbf{b}_{ij}$ are not parameters; rather, they are specified by test analysts to form customized item response models.

Using $\mathbf{a}_{ij}$ and $\mathbf{b}_{ij}$ (or equivalently $\mathbf{a}_{ij}^{*}$ and $\mathbf{b}_{ij}^{*}$) to define the relationship between items and persons allows a general model to be written that includes most of the existing unidimensional Rasch models, such as the Rasch model (Rasch, 1960), the linear logistic test model (Fischer, 1973), the rating scale model (Andrich, 1978), the partial credit model (Masters, 1982), the partial order model (Wilson, 1992), the facets model (Linacre, 1989), the linear rating scale model (Fischer & Parzer, 1991), and the linear partial credit model (Fischer & Pononcy, 1994). The definitions also allow the specification of a range of multidimensional models by imposing linear constraints on the item parameters, such as multidimensional forms of the Rasch model, the rating scale model, the partial credit model, and the linear partial credit model. More important, they allow the specification of the three multidimensional bundle models proposed in this study.

Multidimensional Rasch models have been used to analyze multidimensional items. Kelderman and Rijkes (1994) and Kelderman (1996, 1997) proposed item-scoring functions to describe multidimensionality and analyzed open-ended items about size concepts and multiple-choice items of the Stan-

dard Raven Progressive Matrix and the tests used by the American Society of Clinical Pathologists. Adams et al. (1997) showed that multidimensional Rasch analysis fits the data of a mathematical problem-solving test better than unidimensional Rasch analysis. Hoijtink, Rooks, and Wilmink (1999) investigated the multidimensionality of the Present State Examination (a psychiatric interview), on which some items were designed to measure depression, some were designed to measure anxiety, and the others were designed to measure both depression and anxiety simultaneously. Rost and Carstensen (2002) analyzed a questionnaire that was designed to measure 17 latent traits with 77 items.

Results

Three data sets for the Creativity Development Inventory (Cheng & Wang, 1999) were analyzed. The first data set was obtained from the parallel design with a sample size of 1,210. The second data set was obtained from the sequential design together with a short delay between the two scales and a sample size of 589. The third data set was obtained from the sequential design together with a long delay and a sample size of 352.

The Parallel Design

A total of 1,210 adults and college students in Taiwan took the Creativity Development Inventory. The inventory was designed primarily to identify those characteristics that are important to creativity development but are relatively difficult to possess. The subjects judged how important each characteristic was and how much they possessed it. A subject with a higher score on the first dimension, perceived importance, indicates that he or she values these personal characteristics more. The degree of "valuing" the importance is an individual difference variable because the subjects have different values on the importance of the characteristics, just as people might have different values on other characteristics, such as "wealth." Similarly, a high score on the second dimension, perceived possession, indicates that the subject believes that he or she possesses many of these characteristics. A characteristic with a high rating on the second dimension is one that the subjects deem easy to possess. Characteristics that are important but difficult to possess are the primary tasks to be achieved in cultivating creativity. It is also interesting to know the relationship between perceived importance and perceived possession: Do subjects perceive these characteristics as more important if they possess more of them?

To investigate local item dependence, the original item responses to the 15 items in the two scales were reorganized into 15 bundles, each with nine response categories. The three bundle models were formed. The two latent

variables were assumed to follow a bivariate normal distribution. The computer software ConQuest (Wu, Adams, & Wilson, 1998), being implemented with a marginal maximum likelihood estimation and Bock and Aitkin's (1981) formulation of expectation-maximization algorithm (Dempster, Laird, & Rubin, 1977), was used to estimate the parameters in the three bundle models. The estimation procedure has been described by Adams et al. (1997) and Wu et al. (1998). In addition to ConQuest, the SAS NLMIXED procedure (SAS Institute, Inc., 1999) is an alternative for fitting many common nonlinear and generalized linear mixed models, including the MRCMLM. The reader is referred to Wolfinger and SAS Institute, Inc. (n.d.) for details of the NLMIXED procedure. According to the our experiences in the multidimensional approach, the NLMIXED procedure may take several hours to converge (or sometimes fail to converge), whereas ConQuest takes only a few minutes. Hence, ConQuest was used for all analyses in this study.

In the saturated-dependence bundle model, a total of 123 parameters were estimated, including three parameters for the variance-covariance matrix of the person population distribution (the mean vector was constrained to zero for identification of the parameters), and 15 sets of eight item parameters for the 15 bundles. The variance estimates were 1.08 and 1.57 for the two latent variables, respectively. The covariance estimate was –.06. The correlation between the two latent variables was thus –.05, indicating that the perceived importance of these characteristics was practically unrelated to the perceived possession of them. The saturated-dependence bundle model had a likelihood deviance G^2 (–2 × log likelihood) of 55,135.59. To check whether the model fit the data, the weighted fit Z statistics (Wu et al., 1998) for the 15 bundles are shown in Figure 3. Each bundle had eight Z statistics, one for each parameter. According to the Kolmogorov-Smirnov test of normality, these weighted fit Z statistics appeared to follow approximately the standard normal distribution ($p = .07$), which means that the model fit the data fairly well.

In the uniform-dependence bundle model, a total of 67 parameters were estimated, including 3 parameters for the variance-covariance matrix of the person population distribution, 15 sets of 4 item parameters for the 15 bundles, and 4 dependency parameters. The variance estimates for the two latent variables were 1.08 and 1.56, respectively. The covariance estimate was –.05, and thus the correlation was –.04. They were almost identical to those in the saturated-dependence bundle model. The uniform-dependence bundle model had a likelihood deviance of 55,240.23.

In the independence bundle model, only 63 parameters were estimated, including 3 parameters for the variance-covariance matrix of the person population distribution and 15 sets of 4 item parameters for the 15 bundles. The variance estimates for the two latent variables were 1.11 and 1.57, respectively. The covariance estimate was .41, and hence the correlation was .31, indicating that the more the subjects possessed these characteristics, the more

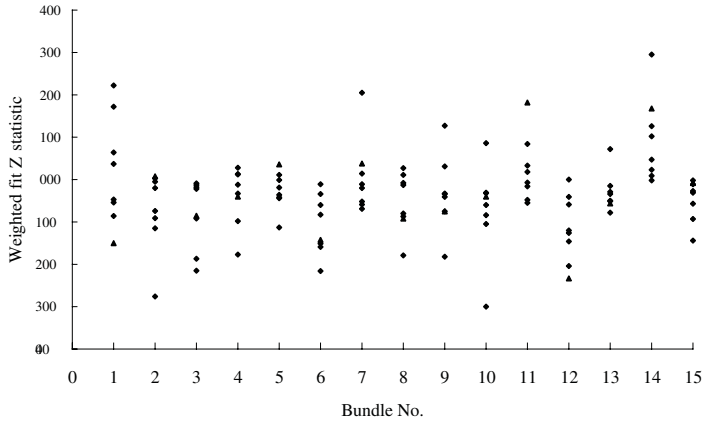

Figure 3. Item fit statistics of the saturated-dependence bundle model under the parallel design.

Table 1
Likelihood Ratio Tests for the Three Nested Bundle Models With the Parallel Design in the First Exemplary Data Set

Model	G^2	Number of Parameters	ΔG^2	*df*	$\Delta G^2/df$
Saturated	55,135.59	123	104.64	56	1.87
Uniform	55,240.23	67	548.51	4	137.13
Independence	55,788.74	63			

they perceived these characteristics as important. The two variance estimates were quite close to those in the saturated-dependence and uniform-dependence bundle models, but the covariance and the correlation were very different from those in the two other models. The independence bundle model had a likelihood deviance of 55,788.74.

Table 1 summarizes the likelihood ratio tests for the three nested models. The saturated-dependence bundle model fit the data statistically better than the uniform-dependence bundle model ($p < .001$), which in turn fit the data statistically better than the independence bundle model ($p < .001$). Because the χ^2 test was very sensitive to large sample sizes, another index was used to account for the model difference: the difference in the two likelihood deviances ΔG^2 was divided by the degrees of freedom. Adding the 4 dependency parameters to the independence bundle model to form the uniform-dependence bundle model reduced the likelihood deviance by 584.51. That is, each dependency parameter accounted for an average of 137.13 in the

deviance, which was very substantial. In contrast, adding another 56 dependency parameters to the uniform-dependence bundle model to form the saturated-dependence bundle model reduced the likelihood deviance to 104.64. Therefore, each of the additional 56 parameters accounted for an average of 1.87 in the deviance, which seemed less substantial. In short, the saturated-dependence bundle model fit the data best, but the uniform-dependence bundle model was promising, which means that the patterns of dependence could be viewed as practically identical across characteristics.

The Sequential Design: A 3-Minute Delay Between Scales

Would the dependence disappear when the sequential design was adopted? To check this, we resampled 589 adults and college students and administered the inventory with the sequential design, in which the scale of perceived importance was printed on one page and the scale of perceived possession on another, as shown in the lower part of Figure 1. The subjects responded to some distracting items between responding to the two scales. These distracting items took approximately 3 minutes. To investigate whether the sequential design washed out the dependence, this new data set was analyzed using the saturated-dependence bundle model and the independence bundle model. Under the saturated-dependence bundle model, the variance estimates for the two latent variables were 0.91 and 1.70, respectively, with a covariance estimate of –.03 and a correlation of –.02, which were very similar to those of the previous saturated-dependence bundle model under the parallel design. It had a likelihood deviance of 27,937.72. Under the independence bundle model, the variance estimates for the two latent variables were 0.94 and 1.70, respectively, with a covariance estimate of .44 and a correlation of .35, which were also quite close to those of the previous independence bundle model under the parallel design. It had a likelihood deviance of 28,289.02. The saturated-dependence bundle model fit the data statistically better than the independence bundle model ($p < .001$), indicating that the dependence still existed under the sequential design with a 3-minute delay. This might have been because the delay of 3 minutes was too short to clear out the dependence. In the following, a much longer delay was used to check whether the dependence would disappear.

The Sequential Design: A 40-Minute Delay Between Scales

A sample of 352 adults and college students took the two scales with a 40-minute delay between them. The saturated-dependence bundle model yielded a likelihood deviance of 21,350.17, variance estimates of 0.85 and

1.33 for the two latent variables, respectively, a covariance estimate of –.06, and a correlation of –.06. The independence bundle model yielded a likelihood deviance of 21,473.32, variance estimates of 0.85 and 1.32 for the two latent variables, respectively, a covariance estimate of .08, and a correlation of .07. The saturated-dependence bundle model fit the data statistically better than the independence bundle model ($p < .001$), meaning that the dependence still existed under the sequential design with a 40-minute delay, yet the change of correlation from –.06 to .07 indicated that it was much reduced.

Summary of the Three Designs

Table 2 lists the likelihood deviance, the correlation between the two latent variables for the saturated-dependence bundle model and the independence bundle model, respectively, the ratio of the two deviances, the difference between the correlations, and the raw score correlations across the three designs. In the parallel design, the dependence was statistically significant ($p < .001$). The correlation changed from .31 to –.05 when the dependence was modeled. The ratio of the two deviances was 98.83%. In the sequential design with a 3-minute delay, the dependence was also statistically significant ($p < .001$). The correlation changed from .30 to –.02 when the dependence was modeled. The ratio of the two deviances was 98.76%. The change in correlation, which could be viewed as an index of the impact of the dependence, in these two designs was remarkable. Considering that the results from the sequential design with a 3-minute delay were almost identical to those from the parallel design, such a short delay did not weaken the dependence.

When the delay was increased to 40 minutes, the correlation changed from .07 to –.06 when the dependence was modeled. The change in correlation was less remarkable than those in the other two designs. The ratio of the two deviances was 99.43%, which was very close to 100% and higher than those of the other two designs. Thus, the saturated-dependence bundle model and the independence bundle model were somewhat practically equivalent. Consequently, the sequential design with a 40-minute delay, though it did not wash the dependence out completely, attenuated it substantially.

Conclusion and Discussion

It has been widely recognized that items with common stimuli might be locally dependent. Unidimensional item response models have been proposed to deal with these kinds of dependence. In this study, we extend the unidimensional methods to multidimensional ones to account for the local item dependence that arises when items across tests are connected by common stimuli. This kind of local item dependence is different from the conventional local item dependence addressed in the literature in that the depend-

Table 2
Summary Statistics for the Saturated-Dependence Bundle Model and the Independence Bundle Model Across the Three Designs

		Saturated		Independence				
Design	*n*	G^2 (1)	*r* (2)	G^2 (3)	*r* (4)	Ratio G^2 (1)/G^2 (3)	Difference *r* (4) – *r* (2)	Raw Score *r*
Parallel	1,210	55,135.59	–.05	55,788.74	.31	98.83%	.36	.28
Sequential: short delay	589	27,937.72	–.02	28,289.02	.30	98.76%	.32	.34
Sequential: long delay	352	21,350.17	–.06	21,473.32	.07	99.43%	.13	.06

ence exists among items across tests rather than within a test and therefore should be modeled through multidimensional item response models. An inventory that was designed to measure perceived importance and perceived possession about creativity-related characteristics is used as an illustration of the dependence. Facing the parallel design, subjects must switch their attention to the characteristics back and forth. To account for the dependence, we adopted a multidimensional bundle method to form the saturated-dependence bundle model, the uniform-dependence bundle model, and the independence bundle model. Using likelihood ratio tests, we compared the three nested models to investigate whether the dependence existed and whether the patterns of dependence across bundles were identical.

From the first exemplary data analysis, one finds that the parallel design indeed caused the dependence and that the patterns of dependence were practically identical across characteristics. If one ignored the dependence, from the results of the independence bundle model, one would claim that perceived importance and perceived possession were moderately and positively correlated ($r = .31$). The correlation was very close to the Pearson correlation between raw scores ($r = .28$). The correspondence is expected because the Rasch score is monotonically related to raw score. In other words, calculating the Pearson correlation between raw scores, which is a standard way to find the correlation, ignores the dependence, as we did in the independence bundle model. Once the dependence was taken into account, from the results of the saturated-dependence bundle model or the uniform-dependence bundle model, one finds that the two latent variables were practically uncorrelated ($r = –.05$). The conclusion of moderate and positive correlation was misleading because the dependence existed but was not taken into account. The conclusion of practically null correlation was more appropriate because the dependence was modeled.

One should be cautious about drawing conclusions of the correlation when the dependence exists. The parallel design is often used mainly for response convenience and economy of space. The dependence is an unexpected disturbance to test users. From a statistical modeling point of view, when a simple model (e.g., the independence bundle model) does not fit the data, one tries using more complicated models (e.g., the saturated-dependence bundle model). In contrast, from a measurement point of view, when unexpected disturbances are found, one may wish to figure out the causes and fix them. The post hoc analytical approach that adopts sophisticated models to deal with unexpected disturbances frees researchers from the hard work of revising research designs and thus is widely used in practice. However, it can cost model parsimony, ease of interpretation, and generalization of the findings. Likewise, in the context of experimental design, when nonrandom errors are found, one may adopt a model that allows correlated errors (e.g., structural equation models for correlated errors), or preferably, one tries to find out the

possible causes and then redesigns the experiment in a well-controlled manner. In this study, we treat the identification of dependence as a warning that asks for a better design (i.e., a sequential design plus a long delay) in which the dependence is minimized.

To reduce the dependence, we conducted the sequential design with a 3-minute delay and collected the second exemplary data set. The dependence was still found and was as intense as that in the parallel design. When the delay was increased to 40 minutes, the dependence still existed but was substantially attenuated. Therefore, the (immediate) sequential design did not necessarily wash out the dependence. The detection of local item dependence should be carried out when common stimuli are used within or between tests, with the parallel design or the sequential design. If dependence is found and its effect is too large to neglect, then either multidimensional dependence bundle models such as those described here should be used to model the data or, preferably, item revision or test redesign should be carried to identify the sources of dependence and the corresponding remedial strategies, guided by conceptual analysis of the psychological constructs underlying the development of the measurement scales.

It is not clear whether the parallel design would definitely cause dependence and how long the delay in the sequential design should be to wash out the dependence completely. These may depend on many contextual factors, such as item characteristics and subject cognitions. For opinion or attitude survey items (e.g., the inventories used in this study), the dependence might be stronger than that for factual items (e.g., "How often do you have a headache? (a) within this week, (b) within this month"). When different scales aim at different specific targets (e.g., "How do you prepare language?" "How do you prepare mathematics?"), the dependence might be weaker than that at vague targets (e.g., "How often do you encounter life stress events?" "How intense are the life stress events?"). Children might be interfered with by the parallel design more strongly and might need a longer delay in the sequential design than adults. More studies are needed to investigate how these factors affect the impact of the dependence and to give clear guidelines for uses of test designs and decisions on delay lengths.

Larger sample sizes are usually needed in IRT than in classical test theory. For bundle analysis, this is especially true, because score patterns rather than individual responses are analyzed. For score patterns that occur rarely, the corresponding parameters would not be acutely calibrated. However, if the detection of dependence rather than the accuracy of parameter estimation is of primary concern, sample sizes need not be very large (the sample size of the third exemplary data set was 352, which was adequate to detect such effects). If sample sizes are small and some score patterns are empty, the corresponding parameters could simply be deleted, which does not affect the generalization of the proposed bundle method. The latent variables are

assumed to be multivariate normal, which is a prerequisite of the Pearson correlation. It is suggested to check the multivariate normality assumption before drawing conclusions about the correlation between latent variables (Johnson, 1998, pp. 67-73).

Local item dependence in this study is treated as fixed effects and is modeled with a set of dependency parameters. As the number of response categories of an item and the number of scales are increased linearly, the number of possible score patterns for an item bundle and the number of the dependency parameters are increased exponentially. For example, if the number of response categories of an item is five rather than three, and the number of scales is three rather than two, then there will be up to 125 ($5 \times 5 \times 5$) possible score patterns for an item bundle so that at most 124 item parameters can be estimated. As a result, in addition to the 12 ($4 + 4 + 4$) step parameters, there will be up to 112 ($124 - 12$) parameters left for the dependence among the 125 response categories. If there are 15 bundles, a total of 1,680 (112×15) dependency parameters will be included under the saturated-dependence bundle model. Modeling so many score patterns with so many dependency parameters is certainly not feasible (sample sizes have to be very large). In such a case, it may be more feasible to treat local item dependence as random effects (Bradlow et al., 1999; Wainer et al., 2000; Wang et al., 2002). This random-effects approach has the strength of dealing with large numbers of response categories of an item and large numbers of scales, because only one single random variable for an item bundle is included. However, this approach, being developed to investigate local item dependence within tests, has not been extended to investigate local item dependence between tests. Further studies are needed to compare implications and applications of the random-effects and fixed-effects approaches for local item dependence between tests.

The underlying item response models used in this study belong to the family of Rasch models. The family of Rasch models has several desirable measurement and psychometric properties, such as observable sufficient statistics for the parameters and relatively small sample size requirements for parameter estimation. The general idea of modeling local item dependence between latent variables can be applied to multidimensional multiparameter item response (MMIR) models (Béguin & Glas, 2001; Bock, Gibbons, & Muraki, 1988; Fraser, 1988; McDonald, 1982; McKinley & Reckase, 1983; Reckase, 1985; Wilson, Wood, & Gibbons, 1991), if the problem of model identification can be resolved properly.

Under the multidimensional three-parameter logistic model for dichotomous items (Hattie, 1981; Reckase, 1985), the probability of a correct response to dichotomous item i for person n is

$$p_{ni} = \gamma_i + (1 - \gamma_{i)} \times \frac{\exp[\boldsymbol{\alpha}_i'(\boldsymbol{\theta}_n - \beta_i \mathbf{1})]}{1 + \exp[\boldsymbol{\alpha}_i'(\boldsymbol{\theta}_n - \beta_i \mathbf{1})]}, \tag{9}$$

where γ_i is the lower asymptote of the item characteristic curve and represents the probability that a person with infinitely low $\boldsymbol{\theta}$ will answer item i correctly; β_i is the difficulty parameter of item i; $\boldsymbol{\alpha}_i$ is a D by 1 vector of the discrimination parameter of item i; and $\mathbf{1}$ is a D by 1 vector of 1's. When there is only one latent variable, equation 9 reduces to the three-parameter logistic model (Birnbaum, 1968). When $\gamma_i = 0$, equation 9 reduces to the multidimensional two-parameter logistic model. Note that $\boldsymbol{\alpha}_i$ and γ_i in equation 9 are both parameters, whereas $\mathbf{a}_{ik}$ and $\mathbf{b}_{ik}$ in equation 7 (the MRCMLM) are not. Also, equation 9 is limited to dichotomous items, whereas equation 7 applies to both dichotomous and polytomous items.

MMIR models are equivalent to exploratory nonlinear factor models. As the factors in exploratory factor analysis are constrained to be orthogonal (i.e., factors are independent) for model identification, so are the latent variables in MMIR models. Under this orthogonality constraint, the derived latent variables are independent, but they are also not readily interpretable because they are usually not consistent with the original test constructs that the test items were designed to measure. Only if the orthogonality constraint is released will the proposed multidimensional bundle method be applicable under MMIR models. To release this constraint, the dimensionality of at least part of the items should be specified, and then one can allow the other items to measure all the latent variables (subject to model identification constraint), which is analogous to confirmatory factor analysis, in which the factor structure is specified rather than discovered from the data. To our knowledge, no commercial computer programs for MMIR models allow users to easily specify the dimensionality of certain items to release the orthogonality constraint. Future studies may aim at developing such computer programs and investigating local item dependence between tests under MMIR models.

References

Adams, R. J., & Wilson, M. R. (1996). Formulating the Rasch model as a mixed coefficients multinomial logit. In G. Englhard & M. Wilson (Eds.), *Objective measurement: Theory into practice* (Vol. 3, pp. 143-166). Norwood, NJ: Ablex.

Adams, R. J., & Wilson, M. R., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1-23.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-573.

Andrich, D. (1985). A latent-trait model for item with response dependences: Implications for test construction and analysis. In S. E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics* (pp. 245-275). New York: Academic Press.

Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and model-fit analysis of multidimensional IRT models. *Psychometrika*, *66*, 541-562.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinees ability. In F. Lord & M. Novick, *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of the EM algorithm. *Psychometrika*, *46*, 443-459.

Bock, R. D., Gibbons, R., & Muraki, E. J. (1988). Full information item factor analysis. *Applied Psychological Measurement*, *12*, 261-280.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153-168.

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavior Statistics*, *22*, 265-289.

Cheng, Y.-Y., & Wang, W.-C. (1999). *Creative thinking and its related factors for teachers with science competition awards*. Research report of the National Science Council, Taiwan.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, *39B*, 1-38.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359-374.

Fischer, G. H., & Parzer, P. (1991). An extension of the rating scale model with an application to the measurement of treatment effects. *Psychometrika*, *56*, 637-651.

Fischer, G. H., & Pononcy, I. (1994). An extension of the partial credit model with an application to the measurement of change. *Psychometrika*, *59*, 177-192.

Fraser, C. (1988). NOHARM: A Fortran program for fitting unidimensional and multidimensional normal ogive models of latent trait theory [Computer software and manual]. Armidale, Australia: University of New England, Centre for Behavioral Studies.

Hattie, J. (1981). *Decision criteria for determining unidimensionality.* Unpublished doctoral dissertation, University of Toronto, Toronto, Canada.

Hoijtink, H., Rooks, G., & Wilmink, F. W. (1999). Confirmatory factor analysis of items with a dichotomous response format using the multidimensional Rasch model. *Psychological Methods*, *4*, 300-314.

Hoskens, M., & De Boeck, P. (1997). A parametric model for local dependence among test items. *Psychological Methods*, *2*, 261-277.

Hoskens, M., & De Boeck, P. (2001). Multidimensional componential item response theory models for polytomous items. *Applied Psychological Measurement*, *25*, 19-37.

Jannarone, R. J. (1986). Conjunctive item response theory kernels. *Psychometrika*, *51*, 357-384.

Jannarone, R. J. (1995). Local dependence: Objectively measures or objectionably abominable? In M. Wilson (Ed.), *Objective measurement: Theory into practice* (Vol. 2, pp. 209-234). Norwood, NJ: Ablex.

Johnson, D. E. (1998). *Applied multivariate methods for data analysts*. Pacific Grove, CA: Brooks/Cole.

Kelderman, H. (1996). Multidimensional Rasch models for partial-credit scoring. *Applied Psychological Measurement*, *20*, 155-168.

Kelderman, H. (1997). Loglinear multidimensional item response models for polytomously scored items. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 287-304). New York: Springer-Verlag.

Kelderman, H., & Rijkes, C. P. M. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika*, *59*, 149-176.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston: Houghton-Mifflin.

Linacre, J. M. (1989). *Many-faceted Rasch measurement*. Chicago: MESA.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

McDonald, R. P. (1982). Linear versus nonlinear models in latent trait theory. *Applied Psychological Measurement*, *6*, 379-396.

McKinley, R. L., & Reckase, M. D. (1983). MAXLOG: A computer program for the estimation of the parameters of a multidimensional logistic model. *Behavior Research Methods and Instrumentation*, *15*, 389-390.

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

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Institute of Educational Research.

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

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

Rosenbaum, P. R. (1988). Item bundles. *Psychometrika*, *53*, 349-359.

Rost, J., & Carstensen, C. H. (2002). Multidimensional Rasch measurement via item component models and faceted designs. *Applied Psychological Measurement*, *26*, 42-56.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *17*, 1-100.

SAS Institute, Inc. (1999). The NLMIXED procedure [Computer software]. Cary, NC: Author.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*, 237-247.

Steinberg, L., & Thissen, D. (1996). Uses of item response theory and the testlet concept in the measurement of psychopathology. *Psychological Methods*, *1*, 81-97.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement*, *26*, 247-260.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Measurement in Education*, *8*, 157-186.

Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*, 185-202.

Wainer, H., & Lukhele, R. (1997). How reliable are TOEFL scores? *Educational and Psychological Measurement*, *57*, 749-766.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement*, *28*, 197-219.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice*, *15*(1), 22-29.

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement*, *37*, 203-220.

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and applications. *Applied Psychological Measurement*, *26*, 109-128.

Wilson, D. T., Wood, R., & Gibbons, R. (1991). TESTFACT: Test scoring, item statistics, and item factor analysis [Computer software and manual]. Mooresville, IN: Scientific Software.

Wilson, M. R. (1988). Detecting and interpreting local item dependence using a family of Rasch models. *Applied Psychological Measurement*, *12*, 353-364.

Wilson, M. R. (1992). The partial order model: An extension of the partial credit model. *Applied Psychological Measurement*, *16*, 309-325.

Wilson, M. R., & Adams, R. J. (1995). Rasch models for item bundles. *Psychometrika*, *60*, 181-198.

Wolfinger, R. D., & SAS Institute, Inc. (n.d.). *Fitting nonlinear mixed models with the new NLMIXED procedure*. Retrieved November 11, 2003, from http://support.sas.com/rnd/app/papers/nlmixedsugi.pdf

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). ConQuest [Computer software and manual]. Camberwell: Australian Council for Educational Research.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three parameter logistic model. *Applied Psychological Measurement*, *8*, 125-145.
Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187-213.