Journal of Educational and Behavioral Statistics
Fall 1997, Vol. 22, No. 3, pp. 265–289

Local Dependence Indexes for Item Pairs Using Item Response Theory

Wen-Hung Chen
ACT

David Thissen
University of North Carolina

Keywords: *IRT, local dependence, Q_3*

Four statistics are proposed for the detection of local dependence (LD) among items analyzed using item response theory. Among them, the X^2 *and* G^2 *LD indexes are of special interest. Simulated data are used to study the distribution and sensitivity of these statistics under the null condition, as well as under conditions in which LD is introduced. The results show that under the null condition of local independence, both the* X^2 *and* G^2 *LD indexes have distributions very similar to the* χ^2 *distribution with 1 degree of freedom. Under the locally dependent conditions, both indexes appear to be sensitive in detecting LD or multidimensionality among items. When compared to* Q_3*, another statistic often used to detect LD, these new statistics are somewhat less powerful for underlying LD, equally powerful for surface LD, and better behaved in the null case.*

Item response theory (IRT) is used for item analysis, test construction, and sometimes test scoring. The validity of these uses requires that the items of a test be locally independent. Local independence means that for a given value of the latent variable θ, the joint probability of correct responses to an item pair is the product of the probabilities of correct responses to the two items,

$$P(X_i = 1, X_j = 1|\theta) = P(X_i = 1|\theta)P(X_j = 1|\theta)\,. \tag{1}$$

However, this may not be true for some pairs of items. Yen (1993) has described several potential causes of local dependence (LD). For example, in a reading comprehension test, the items that follow the same passage may exhibit local dependence. Another example arises when speed is a factor for the test; some of the items at the end may be omitted and thus be locally dependent. In the first case, the local dependence stems from the content of the items, or from the fact that the test is not unidimensional in a psychological sense. In the second case, if the test taker fails to reach item j, he or she will certainly fail to reach item $j + 1$, and so on; these omitted items exhibit local dependence for this

This research project was funded by Law School Admissions Council. The opinions expressed in this report do not necessarily represent those of LSAC.

mechanical reason. Another example, although rare, arises when the test has identical items. Theoretically, a test taker's responses to the identical items should be identical, and the joint probability becomes

$$P(X_i = 1, X_j = 1|\theta_k) = P(X_i = 1|\theta_k) . \tag{2}$$

A pair of items is said to be perfectly locally dependent if every test taker responds identically to both items.

Locally dependent items are redundant; they contain less information than the IRT model would predict. In other words, the test contains fewer items than it appears to. Local dependence also has effects on item parameter estimation. Shown in Table 1 are examples of the effects of local dependence on the parameter estimates using data simulated with the two-parameter logistic (2PL) model. The first example is generated with six locally independent items and 1,000 observations; θ is $N(0, 1)$. The second example is generated in the same manner except that Items 3 and 6 are perfectly locally dependent; that is, the response to Item 6 is identical to the response to Item 3 for every observation. The third example is generated as a two-dimensional test, with the two dimensions correlated .5. The first three items are of one dimension, and the other three items are of the other; this example simulates local dependence within passages of a reading comprehension test. The true item parameters are the same for all three examples and are shown under the heading *true values* in Table 1. The MULTILOG (Thissen, 1991) computer program was used to compute the parameter estimates. If there is local dependence, the parameter estimates may be very different from what they would be if the data were locally independent.

Depending on the purpose of the test, locally dependent items may or may not be desirable. However, if the item parameters are used for item selection, their estimation is affected by local dependence, as the examples above illustrate. In addition, if IRT scaled scores are used, they become inaccurate in the presence of locally dependent items, because the product of the trace lines is not the joint

TABLE 1
Parameter estimates of three sets of simulated data using the 2PL model

			Parameter estimates					
	True values		1st example: locally independent		2nd example: perfect LD pair		3rd example: two-dimensional	
Item	*a*	*b*	*a*	*b*	*a*	*b*	*a*	*b*
1	1.0	−1.0	0.95	−1.05	0.60	−1.41	0.59	−1.37
2	1.0	−0.5	0.95	−0.46	0.51	−0.71	0.61	−0.37
3	1.0	0.0	0.99	−0.01	6.55	0.01	0.84	0.17
4	1.0	0.5	0.89	0.43	0.60	0.63	1.29	0.45
5	1.0	1.0	1.12	0.93	0.49	1.55	0.78	1.37
6	1.0	0.0	0.97	0.04	6.55	0.01	0.76	0.42

likelihood of the item responses. If some degree of local dependence is necessary to meet test specifications, Yen (1993) suggested the construction of testlets to overcome these problems.

Yen (1984) evaluated several statistics as indexes of local dependence: Q_1, variants based on Q_2, and Q_3. The Q_1 statistic (Yen, 1981) was proposed as an overall goodness-of-fit measure for each item of an IRT model and reflects local dependence only indirectly, if at all. The Q_2 statistic (van den Wollenberg, 1982) was designed to be sensitive to local dependence at an aggregate level (over items) for the Rasch model; Yen (1984) extended its application to the 3PL model and suggested the addition of a sign, to construct signed Q_2, to indicate the direction of local dependence. To provide a pairwise index of item dependence, Yen (1984) proposed Q_3, a correlation of residuals from the IRT model, based on ideas suggested by Kingston and Dorans (1982). That is,

$$d_{ik} = u_{ik} - \hat{P}_i(\hat{\theta}_k)\,, \tag{3}$$

and

$$Q_{3ij} = r_{d_i d_j}\,, \tag{4}$$

where u_{ik} is the score of the kth test taker on the ith item. The computation of Q_3 requires a point estimate, $\hat{\theta}$, for each examinee. In the results reported by Yen (1984), and in the implementation of the Q_3 statistic in the computer program PARDUX (Burket, 1993), $\hat{\theta}$ is computed as the mode of the likelihood of the item responses, with no reference to any population distribution. When $\hat{\theta}$ is computed in this way, some response patterns have no finite value. In practice, such response patterns are assigned an arbitrary upper or lower limit; those upper and lower limits have some effect on statistics that are computed using the $\hat{\theta}$ values.

In the results reported here, we use the mode of the posterior over θ as the estimate, $\hat{\theta}$, where the posterior is computed as the product of the response pattern likelihood with an $N(0, 1)$ population distribution. These modal a posteriori (MAP) estimates are more consonant with the maximum marginal likelihood estimation used for the item parameters (in, e.g., BILOG [Mislevy & Bock, 1989] or MULTILOG [Thissen, 1991]) and the integrated posterior methods used in the computation of the LD indexes (indexes of local dependence) proposed here. In addition, no arbitrary solutions are necessary for infinite estimates. However, it should be noted that the values of Q_3 tabulated here are not the same as those previously discussed in the literature (Yen, 1984, 1993).

LD Indexes

The most commonly used method for estimating the parameters of IRT models is maximum marginal likelihood (MML). MML estimation fits the IRT model to the K^m table of response frequencies, where K represents the number of item response categories and m is the number of items. A by-product of this estimation process is the expected frequency for each response pattern, as predicted by the IRT model. The expected and observed frequencies can be assembled into marginal tables for each pair of items. If the IRT model is a good approximation to the data, then the covariation of the observed and expected tables should be approximately the same. If the covariation of the two tables differs more than expected by chance, that is evidence that the IRT model induces more or less dependence than is observed. This concept is similar to the ordinary test of association for contingency tables, except that the expected frequencies come from the IRT model using MML estimation, rather than a row-column loglinear model.

For each pair of items with binary responses the following marginal table can be constructed for the observed frequencies:

		Item *j*	
		0	1
Item *i*	0	O_{11}	O_{12}
	1	O_{21}	O_{22}

In this table O_{pq} is the observed frequency, where 1 and 0 represent the correct and incorrect responses, respectively. The same structure applies to the expected frequencies:

		Item *j*	
		0	1
Item *i*	0	E_{11}	E_{12}
	1	E_{21}	E_{22}

In this table, E_{pq} is the expected frequency that is predicted by the IRT model:

$$E_{pq} = N\int P_i(\theta)^p P_j(\theta)^q [1 - P_i(\theta)]^{(1-p)} [1 - P_j(\theta)]^{(1-q)} f(\theta) d\theta\,, \tag{5}$$

where $P_i(\theta)$ is the trace line for item i, $f(\theta)$ is the population distribution, and N is the number of examinees. The integral is approximated numerically.

Given the goal of constructing LD indexes for pairs of items, it is natural to select existing statistics that serve the purpose of examining the covariation of two-way contingency tables. They should be sensitive to inconsistency between the observed and expected tables. Four potential statistics that are commonly used for testing association in 2 × 2 tables are considered. These four statistics all have the property that they are expected to indicate no association if the table exhibits perfect independence. Their properties have been fully studied under the ordinary test of association situation, that is, for loglinear models.

The first two statistics, the Pearson's X^2 and G^2 statistics, are distributed as χ^2 with degrees of freedom equal to the number of cells minus the number of loglinear parameters; for the test of independence in 2 × 2 tables, the degree of freedom is one. While G^2 and X^2 are asymptotically equivalent (Mood, Graybill, & Boes, 1985, p. 445), they may perform differently for finite samples. In the context of IRT and the LD indexes, an additional parameter, the *slope* or *discrimination* parameter, is used to fit the relationships among items; the slope estimates are derived from the relation of each item with all of the items on the test. When a pair of items is considered in the computation of an LD index, the expected values reflect some degree of dependence, as a function of the items' slope parameters. Thus, when the expected frequencies are predicted from the IRT model, the result may be described as the loss of a fraction of the one degree of freedom for the test of independence. However, if there are more than three items, then the degrees of freedom should be greater than zero, because there are more two-way interactions between all pairs of items than there are slope parameters (Thissen, Bender, Chen, Hayashi, & Wiesen, 1992). The expected values, and simultaneously the degrees of freedom, of these two statistics will be examined using simulated data.

Two other statistics, the standardized ϕ coefficient difference and the standardized log-odds ratio difference, are expected to be distributed $N(0, 1)$. Their expected values are zero under the null hypothesis, that is, local independence. They have an advantage over X^2 and G^2 because they have signs to indicate the direction of association. A positive value of these two statistics indicates greater dependence of the observed frequencies than the IRT model predicts; a negative value indicates less dependence of the observed frequencies than the model predicts. One way to improve the X^2 and G^2 statistics is to incorporate signs in them using a strategy similar to that used by Yen (1984) for the signed Q_2. The signed X^2 and signed G^2 are computed with the signs determined by the sign of $\phi_{\text{obs}} - \phi_{\text{exp}}$.

However, the standardized ϕ coefficient difference and the standardized log-odds ratio difference have a great disadvantage: They are undefined in situations when zero is observed in some of the cells. The standardized ϕ coefficient difference is undefined when both cells of the same row or column have zero observed frequency; the standardized log-odds ratio difference is undefined when any of the cells have zero observed frequency. On the other hand, the contribution of a zero cell to G^2 is defined as zero, so that G^2 is still well defined with empty observed cells.

The definitions of the four statistics are as follows.

Pearson's X^2

Pearson's X^2 is computed as (Bishop, Fienberg, & Holland, 1975, p. 57):

$$X^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}. \tag{6}$$

The Likelihood Ratio G²

The G^2 statistic is computed as (Bishop et al., 1975, p. 58):

$$G^2 = -2 \sum_{i=1}^{2} \sum_{j=1}^{2} O_{ij} \ln\left(\frac{E_{ij}}{O_{ij}}\right). \tag{7}$$

The Standardized φ *Coefficient Difference*

The φ coefficient for the observed frequency table is computed as (Bishop et al., 1975, p. 381):

$$\phi_{\text{obs}} = \frac{O_{11}*O_{22} - O_{12}*O_{21}}{\sqrt{(O_{11} + O_{12})(O_{21} + O_{22})(O_{11} + O_{21})(O_{12} + O_{22})}}. \tag{8}$$

The φ coefficient of the expected frequency table is computed using the same formula, with all O_{pq} replaced by E_{pq}. φ is asymptotically normal with the variance

$$\begin{aligned} \text{Var}(\phi_{\text{obs}}) = \frac{1}{\sum\limits_q \sum\limits_p O_{pq}} \Bigg\{ 1 - \phi^2_{\text{obs}} + (\phi_{\text{obs}} + \frac{1}{2}\phi^3_{\text{obs}}) \\ \left[\frac{(O_{11} + O_{12} - O_{21} - O_{22})(O_{11} + O_{21} - O_{12} - O_{22})}{\sqrt{(O_{11} + O_{12})(O_{21} + O_{22})(O_{11} + O_{21})(O_{12} + O_{22})}}\right] \\ - \frac{3}{4}\phi^2_{\text{obs}} \left[\frac{(O_{11} + O_{12} - O_{21} - O_{22})^2}{(O_{11} + O_{12})(O_{21} + O_{22})} + \frac{(O_{11} + O_{21} - O_{12} - O_{22})^2}{(O_{11} + O_{21})(O_{12} + O_{22})}\right]. \end{aligned} \tag{9}$$

The standardized φ coefficient difference is then

$$\phi_{\text{diff}} = \frac{\phi_{\text{obs}} - \phi_{\text{exp}}}{\sqrt{\text{Var}(\phi_{\text{obs}})}}. \tag{10}$$

The Standardized Log-Odds Ratio Difference (τ)

The log-odds ratio of the observed table is computed as (Haberman, 1978, p. 105):

$$\tau_{\text{obs}} = \ln\left(\frac{O_{11}*O_{22}}{O_{12}*O_{21}}\right). \tag{11}$$

The log-odds ratio of the expected table is computed in the same manner, when O_{pq} is replaced by E_{pq}. The standardized log-odds ratio difference is equal to

$$\frac{\tau_{\text{obs}} - \tau_{\text{exp}}}{\sqrt{\sum\limits_p \sum\limits_q {}^1/O_{pq}}}. \tag{12}$$

The denominator is the standard deviation of the log-odds ratio statistic.

Table 2 shows the observed and expected frequencies based on the 3PL IRT model for two pairs of items from an admissions test, along with the values of the four statistics above and Q_3 associated with these two pairs. The statistics suggest that the first pair of items, Item Pair A, are locally independent; for the second pair of items, Item Pair B, local dependence is strongly indicated by these statistics.

Models of Local Dependence

Underlying Local Dependence

The first type of local dependence to be considered is called *underlying local dependence* (ULD; Thissen et al., 1992). This model assumes that there is a separate trait that is common to each set of locally dependent items but is not common to the rest of the items in the test. The θ associated with each item is determined by some weights, θ^*_1, and the separate traits θ^*_k. The relationship of θ and the underlying parameters θ^*_k for two pairs of items is

$$\begin{bmatrix} \theta_{\text{item}_1} \\ \theta_{\text{item}_2} \\ \theta_{\text{item}_3} \\ \theta_{\text{item}_4} \end{bmatrix} = \begin{bmatrix} wt_1 & wt_{\text{item}_{12}} & 0 \\ wt_2 & wt_{\text{item}_{22}} & 0 \\ wt_3 & 0 & wt_{\text{item}_{33}} \\ wt_4 & 0 & wt_{\text{item}_{43}} \end{bmatrix} \begin{bmatrix} \theta^*_1 \\ \theta^*_2 \\ \theta^*_3 \end{bmatrix} \tag{13}$$

where wt_i and $wt_{\text{item}_{ik}}$ are the weights; this gives the result that Items 1 and 2 are locally dependent, as are Items 3 and 4. The weights for the noncommon second traits are always zero. θ and θ^*_k are defined so that they are distributed $N(0, 1)$, so the sum of squares of each row of the weight matrix should be one. Each θ is the linear combination of θ^*_1 and one of the θ^*_ks. If all $wt_{\text{item}_{ik}}$ equal zero, then all wt_i equal one, and θ equals θ^*_1, which gives locally independent data. In that case, the test is said to be unidimensional.

TABLE 2
The observed and expected frequencies based on the 3PL IRT model for two pairs of admissions test items with their corresponding LD indexes and Q_3

	Item Pair A			Item Pair B	
	0	1		0	1
0	Obs: 599 Exp: 599	Obs: 266 Exp: 261	0	Obs: 654 Exp: 522	Obs: 203 Exp: 331
1	Obs: 536 Exp: 524	Obs: 380 Exp: 397	1	Obs: 346 Exp: 473	Obs: 539 Exp: 416
	G^2:	1.10		G^2:	159.19
	X^2:	1.09		X^2:	153.28
	ϕ_{diff}:	−0.88		ϕ_{diff}:	13.42
	τ_{diff}:	−0.87		τ_{diff}:	12.15
	Q_3:	−0.01		Q_3:	0.33

If all wt_i equal zero, then all $wt_{\text{item}_{ik}}$ must be one, and the test is multidimensional with uncorrelated dimensions. In fact, items with this kind of structure should be considered as multiple tests, and IRT analysis should be applied separately to each set of items. The *i*th and *j*th items of a set are locally dependent when wt_i, wt_j, $wt_{\text{item}_{ik}}$, and $wt_{\text{item}_{jk}}$ are nonzero. If that is true also for other sets of items then the test is said to be multidimensional with the dimensions correlated. This is intended to be similar to a reading comprehension test with a set of items following each passage. This last situation is the focus of the ULD model.

When estimating parameters with correlated multidimensional data using a unidimensional IRT model, a combination of the underlying dimensions becomes the latent variable (Yen, 1984). Under such circumstances, both positive and negative local dependencies will appear among pairs of items. Yen observed that positive local dependence exists within sets of items that measure the same second trait, and negative local dependence exists between sets of items that measure different traits. Thus, the four potential LD indexes should have positive signs when a pair of items is influenced by a single second trait, as are Items 1 and 2, and Items 3 and 4, in (13). They should have negative signs when a pair of items is influenced by different traits, as are Items 1 and 3, and Items 1 and 4, in (13). Analyzing data simulated with the ULD model enables us to examine this property.

Surface Local Dependence

The second type of local dependence to be modeled is called *surface local dependence* (SLD; Thissen et al., 1992). Examples of this type of local dependence arise on a long test where some of the items at the end are omitted, or in a test with similar items. The idea is that a pair of items are so similar (in content or in location in the test) that the test taker responds identically to the second item without the underlying processing implied by the IRT model (Thissen et al., 1992). For a pair of items, a model for such a process is the following:

$$\begin{aligned}
&\text{With probability } 1 - \pi_{\text{LD}}\text{, the IRT model:}\\
&\qquad \text{Response to Item 2} = \begin{cases} 1, \text{ with } P(X_2 = 1|\theta) \\ 0, \text{ with } P(X_2 = 0|\theta). \end{cases}\\
&\text{With probability } \pi_{\text{LD}}\text{:}\\
&\qquad \text{Response to Item 2} = \begin{cases} 1, \text{ if } X_1 = 1 \\ 0, \text{ if } X_1 = 0. \end{cases}
\end{aligned} \tag{14}$$

In the model, π_{LD} is the probability that the test taker will respond to the second item in the same way as to the first item without regard to θ. If π_{LD} equals zero, then there is no local dependence, and the model is the ordinary IRT model. If π_{LD} equals one, then the two items are said to be perfectly locally dependent. This model is a variant of the model for item dependency suggested by Ackerman and Spray (1986). The Ackerman and Spray model depends on two parameters, which they call α and β; those parameters may be set to induce asymmetric LD for an item pair. The relation between the two models is that $\pi_{\text{LD}} = (1 - \alpha) = (1 - \beta)$; the SLD model used here always induces symmetric LD.

It is easy to imagine that as π_{LD} approaches one, the increasing local dependence has more of an effect on the parameter estimates. When a pair of items exhibits perfect local dependence, the maximum likelihood estimate of the slope parameters may be infinite (Thissen et al., 1992). Influenced by the extreme covariation of this pair of items, the local dependence becomes the underlying dimension of the IRT model. As a result, the slope estimates of the remaining items are reduced to reflect only their covariance with this dimension, as shown in Table 1.

The proposed LD indexes should be able to identify such a pair. However, two questions need to be answered: First, how strong is this effect when the ratio between the number of locally independent items and number of locally dependent items becomes large? The second question is how strong the effects of the local dependence may be when the value of π_{LD} is between 0 and 1. To what extent can the proposed LD indexes detect less-than-perfect local dependence? These questions will be examined using simulated SLD data with varying values of π_{LD} and varying numbers of items. Parallel questions will be considered regarding the performance of the LD indexes in the presence of ULD. As a prerequisite to all of this, the null distributions of the new LD indexes and Q_3 will be examined. Finally, comparisons will be drawn between the performance of the new LD indexes and Q_3.

Method

Data Simulation

Three types of data were generated to study the properties of the four statistics, using the computer program GEN8 (Chen, 1994). The item parameters were randomly generated from three distributions that closely resemble the distribution of item parameters obtained in practice: The slopes were sampled from a log-normal(0, 0.5) distribution, the threshold parameters from an $N(0, 1.5)$ distribution, and the guessing parameters from a logit-normal(-1.1, 0.5) distribution. θ for each observation was randomly sampled from the $N(0, 1)$ distribution. We generated locally independent items for the 2PL and 3PL models, for completely separately generated tests with 10, 40, and 80 items and 1,000 examinees. These data were used to determine the distributions and expected values of these statistics under the null hypothesis.

For the ULD model, we first simulated data for 40 items with one general factor and four specific factors. Twenty of these 40 items loaded on the general factor only. That is, for these 20 items the wt_i were one and $wt_{\text{item}_{ij}}$ were zero. The remaining 20 items also loaded on the four specific factors, in addition to the general factor, with 5 items loading on each factor and each item loading on one and only one specific factor. The $wt_{\text{item}_{ij}}$ were generated from the $N(0.71,$

0.1) distribution, truncated to the interval [0.4, 1.0]; thus, the average interitem correlation for the underlying variable for items on the same factor derived from the specific factors is $0.71^2 = 0.5$. The θ_1^* and θ_k^* were generated in the same manner as θ. These simulations generated strong multidimensionality; the average weights of the items on the general and specific factors are equal.

In addition, a second set of simulations was generated with two specific factors, and different numbers of items. In this set, the $wt_{\text{item}_{ij}}$ were generated from normal distributions with different means. The weights of the first specific factor were generated from the $N(0.6, 0.1)$ distribution, truncated to the interval [0.3, 0.9]. The weights of the second specific factor were generated from the $N(0.4, 0.1)$ distribution, truncated to the interval [0.1, 0.7]. This simulation may be considered weakly multidimensional, because the average weight on the general factor is larger than the average weights of the two specific factors. Table 3 lists the conditions for simulating ULD data sets.

To simulate item responses for the SLD model, the parameter π_{LD} of any locally dependent item pair was either 0.3, 0.5, or 0.8, to examine the sensitivity of the LD indexes in situations where the local dependence is low, moderate, or strong. At this time, we consider items only in pairs. Although (14) can be extended to the situation with more than two items, the formula becomes complicated, and eventually a stochastic model (e.g., random walk) may be involved. We used a 1/10 ratio in this study; that is, we generated 1 pair of surface locally dependent items for every 10 items. Table 4 lists the conditions for simulating SLD data sets; all sample sizes were 1,000.

Altogether, we simulated data for 24 different conditions: 6 conditions for the locally independent condition, 8 for the ULD models, and 10 for the SLD model. There were 100 replications for each condition, and each replication contained 1,000 observations.

TABLE 3
Conditions for simulating underlying locally dependent data

IRT model	No. of items	No. of items on each factor	Weight distribution of loadings on specific factors
		Strong multidimensionality	
2PL	40	20, 5, 5, 5, 5	$N(0.71, 0.1)$
3PL	40	20, 5, 5, 5, 5	$N(0.71, 0.1)$
		Weak multidimensionality	
2PL	10	5, 5	$N(0.4, 0.1)$, $N(0.6, 0.1)$
	40	20, 20	$N(0.4, 0.1)$, $N(0.6, 0.1)$
	80	40, 40	$N(0.4, 0.1)$, $N(0.6, 0.1)$
3PL	10	5, 5	$N(0.4, 0.1)$, $N(0.6, 0.1)$
	40	20, 20	$N(0.4, 0.1)$, $N(0.6, 0.1)$
	80	40, 40	$N(0.4, 0.1)$, $N(0.6, 0.1)$

TABLE 4
Conditions for simulating surface locally dependent data

IRT model	No. of items	π_{LD}	No. of LD item pairs
2PL	10	0.5	1
	40	0.3	4
	40	0.5	4
	40	0.8	4
	80	0.5	8
3PL	10	0.5	1
	40	0.3	4
	40	0.5	4
	40	0.8	4
	80	0.5	8

Parameter Estimation

The MULTILOG (Thissen, 1991) computer program was used to estimate the parameters of the IRT models. Three procedures were applied in every IRT analysis in order to obtain a better fit of the data. First, the number of EM cycles was increased from the default 25 to 100. Second, prior distributions were used for all item parameters; the prior distributions were chosen from among those available in MULTILOG to approximate the generating distributions. For the 2PL model, the prior used for the slope parameter is the $N(1.1, 0.6)$ distribution; the prior used for the threshold parameter is the $N(0.0, 1.5)$ distribution. For the 3PL model, the prior used for the slope parameter is the $N(1.9, 1.0)$ distribution; the prior used for the threshold parameter is the same as that for 2PL model; and the prior used for the lower asymptote parameter is the logit-normal $(-1.1, 0.5)$ distribution. Finally, the number of quadrature points was increased from the program's default to obtain more precise estimation of the item parameters (Thissen & Chen, 1993), avoiding error due to inadequate accuracy in the numerical integration. For the most simple condition—that is, the 2PL model with 10 items—10-point quadrature was used. Nineteen-point quadrature was used for the 3PL model with 10 items and for either model with 40 items, and 37 quadrature points were used for either model with 80 items.

The Computation of the LD Indexes and Q_3

The IRT__LD (Chen, 1993) computer program was used to compute the proposed LD indexes and Q_3. The program used the parameter estimates from MULTILOG (Thissen, 1991) to compute the expected frequencies. It then used the observed and expected frequencies to compute the LD indexes. For the Q_3 statistics, it used the same item parameter estimates to estimate the individual $\hat{\theta}$, then computed Q_3 as described in (3) and (4).

The means and standard deviations of the proposed X^2 and G^2 LD indexes were computed for the locally independent data, pooling all replications. The means obtained from the locally independent condition enable us to examine the distribution of each statistic under the null hypothesis.

For the ULD model, the proportions of the G^2 LD index that exceed the nominal levels ($\alpha = 0.05$ and 0.01) for $\chi^2(1)$ were computed separately based on the factor(s) represented in the item pairs. For the SLD model, the proportions of the G^2 LD index that exceed the nominal levels ($\alpha = 0.05$ and 0.01) for $\chi^2(1)$ were also computed separately for the locally independent item pairs and locally dependent item pairs.

The Q_3 statistics were computed for the locally independent condition and the ULD model under strong multidimensionality, as well as for the SLD model with 40 items and various values of π_{LD}, namely, 0.05, 0.1, 0.3, and 0.5. These enable us to compare Q_3 and the proposed LD indexes for their Type I error rates and their power.

Results

The Null Distribution of the LD Indexes

Table 5 shows the means and standard deviations for the X^2 LD index and G^2 LD index for the 2PL and 3PL models for the locally independent case. This table shows that there is almost no difference between the X^2 and G^2 LD indexes with the 2PL model; but the X^2 LD index has slightly smaller means and standard deviations with the 3PL model. There is a tendency for the means and standard deviations to increase as the number of items increases.

As suggested earlier, these means should represent the expected values of a χ^2 distribution with degrees of freedom less than one. The data in Table 5 confirm this speculation: Most of the means are less than one, and the standard deviations are less than 1.4. When the number of items increases, the ratio of the number of item × item tables to the number of item slope parameters becomes large, and the degrees of freedom approaches one (Thissen & Chen, 1993). This tendency is obvious in these tables. Three of the means are above one, but not very different from one; those may be due to random sampling. Alternatively, given the fact that the means greater than one all appear using the 3PL model, it may be that the estimation of the lower asymptote has some modest biasing effect on X^2 and G^2.

TABLE 5
Means and standard deviations of the X² *and* G² *LD indexes for 100 replications, simulated locally independent data, for the 2PL and 3PL models, with 1,000 observations*

No. of items	Means				Standard deviations			
	2PL		3PL		2PL		3PL	
	X^2	G^2	X^2	G^2	X^2	G^2	X^2	G^2
10	0.69	0.69	0.84	0.90	0.94	0.98	1.08	1.21
40	0.81	0.83	0.97	1.03	1.11	1.11	1.17	1.29
80	0.89	0.90	1.06	1.13	1.15	1.15	1.20	1.31

Table 6 shows the means and standard deviations of the standardized ϕ coefficient difference (ϕ_{diff}) and the standardized log-odds ratio difference (τ_{diff}). In some cases, when an observed frequency is zero, these two statistics are undefined. Therefore, some item pairs are omitted from the computations for Table 6; this may account for the apparent negative bias in the statistics. Due to the apparent negative bias in these statistics, and the fact that they become undefined in some cases, the useful feature of these two statistics appears to be their signs, indicating the direction of association. Thus, for most part, these two statistics will be omitted from further discussion.

The Type I Error Rates of the X^2 *and* G^2 *LD Indexes*

One possible use of an LD index would involve selecting critical values and flagging item pairs as locally dependent if the obtained LD index exceeds the critical values. For X^2 and G^2, suppose we select 3.84 for a 5% or 6.63 for a 1% Type I error rate on a per-item-pair basis, because those are the nominal αs of 0.05 and 0.01 for the χ^2 distribution with one degree of freedom. What then is the true α level for the LD indexes, under the nominal 5% or 1% levels?

Table 7 shows the percentages of the X^2 and G^2 LD indexes that exceed 3.84 and 6.63. These observed percentages suggest that for the nominal 5% level, the α varies between 1.6% and 3.6% for the X^2 statistic and between 1.8% and 5% for the G^2 statistic. For the nominal 1% level, the α varies between 0.2% and 0.5% for the X^2 statistic and between 0.2% and 0.7% for the G^2 statistic. The α consistently increases as the number of items increases, and from the 2PL model to the 3PL model, but never exceeds the nominal level of 5% or 1%. These results correspond with the fact that the means are less than 1.0, and the standard deviations are less than 1.4. The X^2 and G^2 LD indexes appear to be distributed very nearly as a χ^2 distribution with degrees of freedom slightly less than one in the null condition. Using the critical values of 3.84 and 6.63 of the χ^2 distribution with one degree of freedom is slightly conservative.

Table 8 shows the observed 95th and 99th percentiles of the X^2 and G^2 LD indexes under the null condition. They suggest that the true 0.05 critical value may be located around 3.38 and 3.84 for X^2 and G^2, respectively, when the 3PL

TABLE 6

Means and standard deviations of the standardized ϕ *coefficient difference and standardized log-odds ratio difference LD indexes for 100 replications, simulated locally independent data, for the 2PL and 3PL models, with 1,000 observations*

No. of items	Means				Standard deviations			
	2PL		3PL		2PL		3PL	
	ϕ_{diff}	τ_{diff}	ϕ_{diff}	τ_{diff}	ϕ_{diff}	τ_{diff}	ϕ_{diff}	τ_{diff}
10	−0.07	−0.07	−0.06	−0.04	1.13	0.80	0.96	0.86
40	−0.01	−0.02	−0.06	−0.02	0.99	0.87	1.07	0.90
80	−0.05	−0.01	−0.02	−0.01	1.01	0.88	1.08	0.91

TABLE 7
Percentages of the X^2 *and* G^2 *indexes greater than 3.84 and 6.63 for 100 replications, simulated locally independent data, for the 2PL and 3PL models, with 1,000 observations*

No. of	2PL		3PL	
items	X^2	G^2	X^2	G^2
		Percentages greater than 3.84		
10	1.6	1.8	2.6	4.0
40	2.6	2.7	3.2	4.7
80	2.9	3.1	3.6	5.0
		Percentages greater than 6.63		
10	0.2	0.2	0.4	0.4
40	0.4	0.4	0.5	0.7
80	0.4	0.5	0.4	0.7

TABLE 8
95th and 99th percentiles of X^2 *and* G^2 *LD indexes for 100 replications, simulated locally independent data, for the 2PL and 3PL models, with 1,000 observations*

No. of	2PL		3PL	
items	X^2	G^2	X^2	G^2
		95th percentiles		
10	2.47	2.53	3.01	3.48
40	2.95	3.04	3.24	3.73
80	3.11	3.19	3.38	3.84
		99th percentiles		
10	4.52	4.82	5.18	5.55
40	5.18	5.32	5.55	5.95
80	5.36	5.49	5.65	6.07

model is used to analyze data generated with that model. The true 0.01 critical values are about 5.65 and 6.07 for X^2 and G^2, respectively, under the same circumstances.

The Power of the G^2 *LD Index When There Is Local Dependence*

Table 9 shows the percentages of the G^2 LD index that exceed 3.84 and 6.63 under the ULD model. When the average weights on the general and specific factors are equal, for the 2PL model about 40% of the locally dependent item pairs exceed the nominal level $\alpha = 0.05$ for $\chi^2(1)$, and about 25% for $\alpha = 0.01$. For 3PL model, 29–37% of locally dependent item pairs exceed $\alpha = 0.05$, and 17–23% exceed $\alpha = 0.01$. For both models, these percentages never exceed 5 or 1% for the locally independent item pairs (pairs of items that load only on the general factor, and pairs of items that do not load on the same factors). Thus, we have moderate power to detect locally dependent item pairs, with the Type I

TABLE 9

Percentages of the G^2 *LD index greater than 3.84 and 6.63 for 100 replications, simulated underlying locally dependent data*

	% > 3.84		% > 6.63	
Category of item pair	2PL	3PL	2PL	3PL
First specific factor	40.2	36.7	23.3	20.2
Second specific factor	37.1	31.1	24.2	18.1
Third specific factor	40.1	28.8	27.2	17.0
Fourth specific factor	39.5	37.1	23.6	23.0
General factor	2.5	4.6	0.4	0.7
Cross-factors	5.0	4.7	0.9	0.7

Note. No. of items = 40, N = 1,000, weight on specific factor is N(0.71, 0.1). There are 5 items on each specific factor; the remaining 20 items are loaded on the general factor only.

error rate nicely contained below the nominal α level, when the average weights on the general and the specific factors are equal.

However, when the multidimensionality is not so obvious—that is, when the average weights for the specific factors are lower than that of the general factor—the result is that a lower percentage of item pairs exceed the nominal α level. Nonetheless, when compared with the percentage under the null condition, even this low percentage suggests the existence of local dependence (the percentages are 2–3 times higher than in the null condition). Table 10 shows the percentages of the G^2 LD index that exceed 3.84 and 6.63 under the weakly multidimensional condition, where it exhibits less power.

Table 11 shows the percentages of the G^2 LD index that exceed the 3.84 and 6.63 critical values of the $\chi^2(1)$ distribution, for several types of item pairs. *LI*

TABLE 10

Percentages of the G^2 *LD index greater than 3.84 and 6.63 for 100 replications, simulated underlying locally dependent data*

	% > 3.84			% > 6.63		
	10 items	40 items	80 items	10 items	40 items	80 items
		2PL model				
First specific factor	9.2	7.0	7.2	1.8	1.9	1.9
Second specific factor	5.7	6.2	6.9	1.4	1.4	1.7
Cross-factor	4.4	8.5	10.8	0.7	2.0	2.8
		3PL model				
First specific factor	7.2	8.8	9.5	2.0	2.1	2.4
Second specific factor	5.4	9.2	8.5	1.4	2.1	1.9
Cross-factor	5.1	7.7	8.3	0.6	1.4	1.6

Note. N = 1,000, weights on specific factors are N(0.6, 0.1) and N(0.4, 0.1).

TABLE 11
Percentages of the G^2 LD index greater than 3.84 or 6.63 for 100 replications, simulated surface locally dependent data, π_{LD} = 0.5

	2PL model		3PL model	
Category of item pair	% > 3.84	% > 6.63	% > 3.84	% > 6.63
	No. of items = 10			
LI pairs	9.9	3.6	13.8	6.0
Mixed LI & SLD pairs	22.4	7.4	19.3	6.4
Mixed SLD pairs				
SLD pairs	100.0	100.0	100.0	99.0
	No. of items = 40			
LI pairs	3.0	0.4	5.2	0.6
Mixed LI & SLD pairs	3.6	0.6	5.0	0.8
Mixed SLD pairs	7.4	1.1	6.3	0.9
SLD pairs	99.8	99.5	99.8	98.5
	No. of items = 80			
LI pairs	3.2	0.5	5.0	0.8
Mixed LI & SLD pairs	3.2	0.5	4.9	0.7
Mixed SLD pairs	4.8	1.6	6.4	1.8
SLD pairs	87.8	87.3	88.3	86.5

pairs are made up of locally independent items, and *SLD pairs* are pairs of items for which π_{LD} for that item pair is greater than zero. A *mixed LI and SLD pair* includes one item that is locally independent and one item that is locally dependent on some other item outside the pair. A *mixed SLD pair* includes two items that are both locally dependent on some other items outside the pair. In summary, the power is always slightly lower for the 3PL model. The power for the G^2 LD index is about 1.0 with 10 items and decreases as the number of items increases. The power is smallest when π_{LD} = 0.3, as shown in Table 12, but the difference in power across values of π_{LD} is minor compared with the difference in power for different numbers of items. For 40 items, with four locally dependent item pairs, the power is 1.0 with π_{LD} = 0.8 and drops to 0.97 with π_{LD} = 0.3, when considering data generated and analyzed using a 3PL model. In all conditions, the proportion of the G^2 LD index that exceed the nominal 5% or 1% value is never less than 86% (at 80 items with π_{LD} = 0.5).

The Null Distribution of Q_3

Q_3 is the Pearson product-moment correlation of a set of residuals from the IRT model. Scatterplots of those residuals for two pairs of locally dependent items are shown in Figures 1 and 2. The scatterplots are divided into four quadrants, because there are four classes of residuals for a pair of items. When both items are correct, both residuals are computed as the difference between 1 and the trace line, and both residuals are between 0 and 1. When both items are

TABLE 12

Percentages of the G^2 *LD index greater than 3.84 or 6.63 for 100 replications, simulated surface locally dependent data, for 40 items*

	2PL model		3PL model	
Category of item pair	% > 3.84	% > 6.63	% > 3.84	% > 6.63
	$\pi_{LD} = 0.3$			
LI pairs	2.9	0.4	4.9	0.7
Mixed LI & SLD pairs	3.2	0.5	4.5	0.7
Mixed SLD pairs	3.5	0.6	4.4	0.6
SLD pairs	98.5	98.0	98.5	97.3
	$\pi_{LD} = 0.5$			
LI pairs	3.0	0.4	5.2	0.6
Mixed LI & SLD pairs	3.6	0.6	5.0	0.8
Mixed SLD pairs	7.4	1.1	6.3	0.9
SLD pairs	99.8	99.5	99.8	98.5
	$\pi_{LD} = 0.8$			
LI pairs	3.1	0.5	5.2	0.7
Mixed LI & SLD pairs	4.2	0.7	5.5	0.8
Mixed SLD pairs	15.0	4.5	12.5	3.7
SLD pairs	100.0	100.0	100.0	100.0

incorrect, both residuals are computed as the difference between 0 and the trace line, and both residuals are between 0 and −1. When one item is correct and the other item is incorrect, the points lie in the (+, −) or (−, +) quadrant.

The prominent feature of these scatterplots is that the residuals lie along four curved lines, one line in each of the four quadrants, because within each quadrant the location of the points is determined entirely by the locations of the two trace lines. Figure 1 shows the residuals for a pair of items in which the first is of moderate difficulty ($b = -0.18$), while the second item is more difficult ($b = 1.55$); Figure 2 shows the residuals for a pair of items in which the first is more difficult ($b = 0.55$), while the second item is very easy ($b = -3.43$). For any pair of items, the residuals always lie along four curved lines. The locations (and degree of curvature) of those lines vary depending on the difficulty and discrimination of the items in the pair.

Figures 1 and 2 also show the number of examinees in each of the four quadrants. All of the LD indexes consider only those four numbers in determining the degree of local dependence between a pair of items: If those four frequencies are as expected from the IRT model, all of the LD indexes take on small values, and if those four frequencies differ from those predicted from the IRT model, local dependence is indicated.

In contrast, Q_3 has at least four components to its value, and those four components can work in different directions:

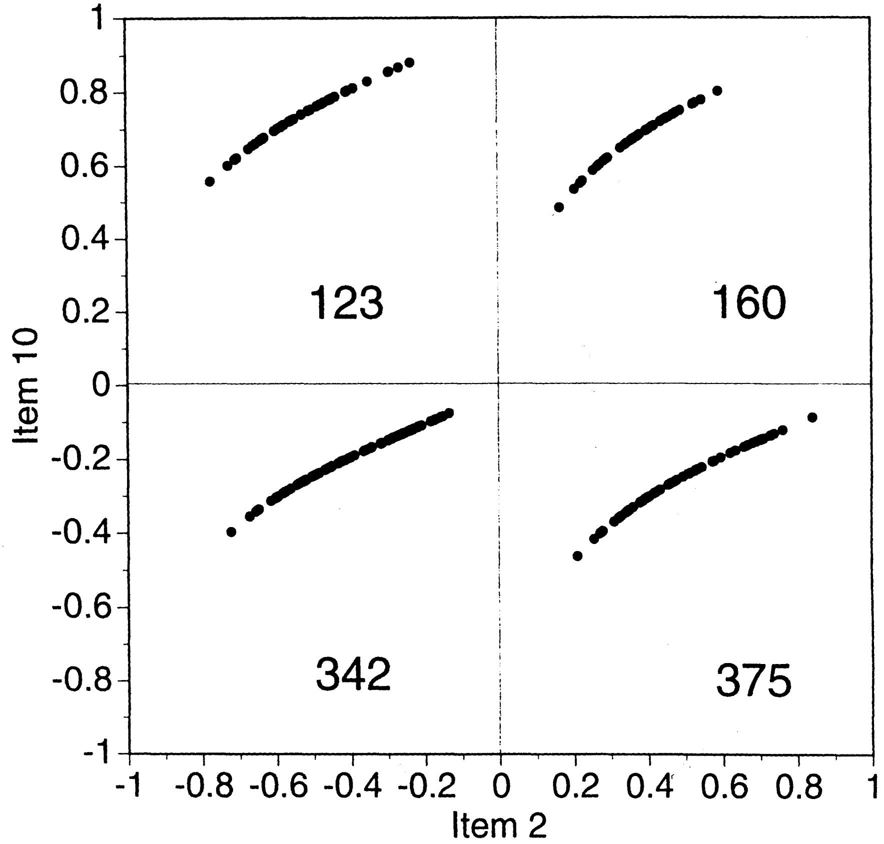

FIGURE 1. *Residual plot of an item pair from a simulated surface locally dependent 2PL data set of 10 items*

Note. The G^2 LD index for this item pair is 4.41, and Q_3 is -0.12. The estimated slope parameters are 0.89 and 0.65, and the threshold parameters are -0.18 and 1.55 for Items 2 and 10, respectively.

(1) As observed by Yen (1984, 1993), Q_3 has a tendency to be slightly negative in the null case because $\hat{\theta}$ is estimated from these same data. The estimation of $\hat{\theta}$ is such that the (weighted) sum (over items) of the residuals is zero. As a result, a slight negative autocorrelation is induced in the residuals, with the result that the expected value of Q_3 is slightly negative.

(2) On the other hand, the fact that all of the trace lines are monotonically increasing means that the four curves along which the residuals fall are all somewhat positively oriented (see Figures 1 and 2). This creates a tendency for Q_3 to be positive. In the null case, this effect does not overcome the negative autocorrelation, so the expected value does not become positive. However, it does introduce some degree of skewness in the distribution of the statistic that is exacerbated for some kinds of local dependence.

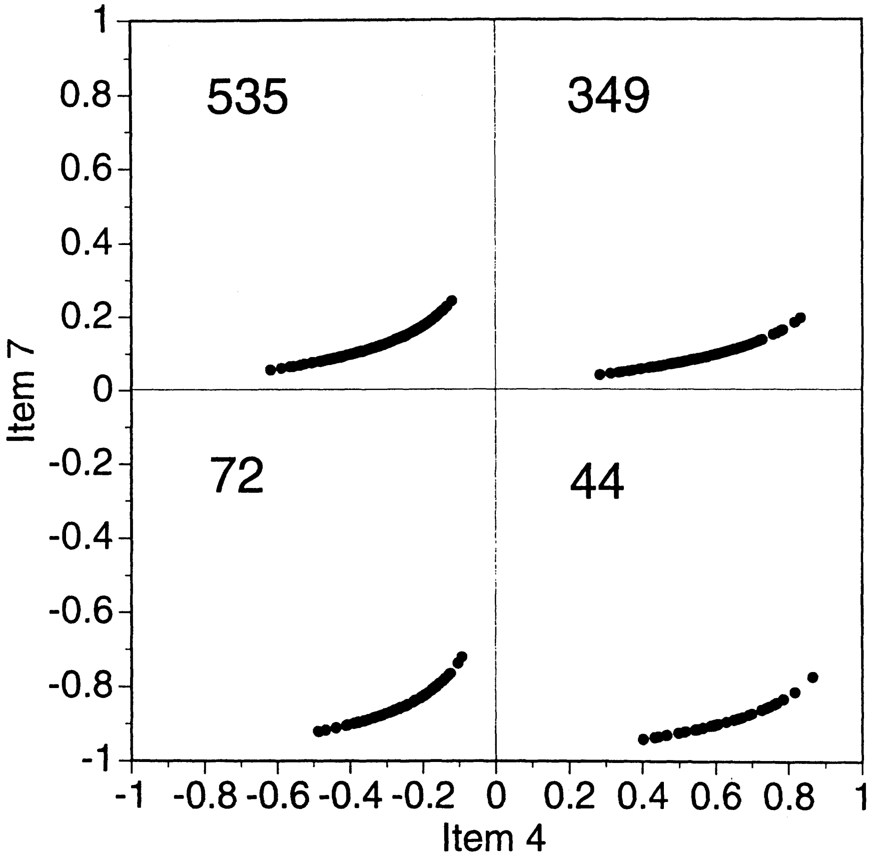

FIGURE 2. ***Residual plot of an item pair from a simulated underlying locally dependent 2PL data set of 10 items***

Note. The G^2 LD index for this item pair is 4.10, and Q_3 is -0.094. The estimated slope parameters are 0.92 and 0.63, and the threshold parameters are 0.55 and -3.43 for Items 4 and 7, respectively.

(3) The difficulty, discrimination, and lower asymptote values of the two items in the pair also have an effect on the value of Q_3, due to their effects on the location of the four curves in the residual space.

(4) Finally, the relative number of residuals in each of the four quadrants has an effect on the value of Q_3; this last contributor is the only aspect in common with the integrated LD indexes described above.

Yen (1984, 1993) suggested that in the null case the mean of the sampling distribution of Q_3 ought to be about $1/(n - 1)$ and that the variance of the Fisher z transform should be about $1/(n - 3)$. These are normal theory values and would be correct if the data being correlated arose from a bivariate Gaussian distribution. However, Figures 1 and 2 illustrate the fact that the residuals being

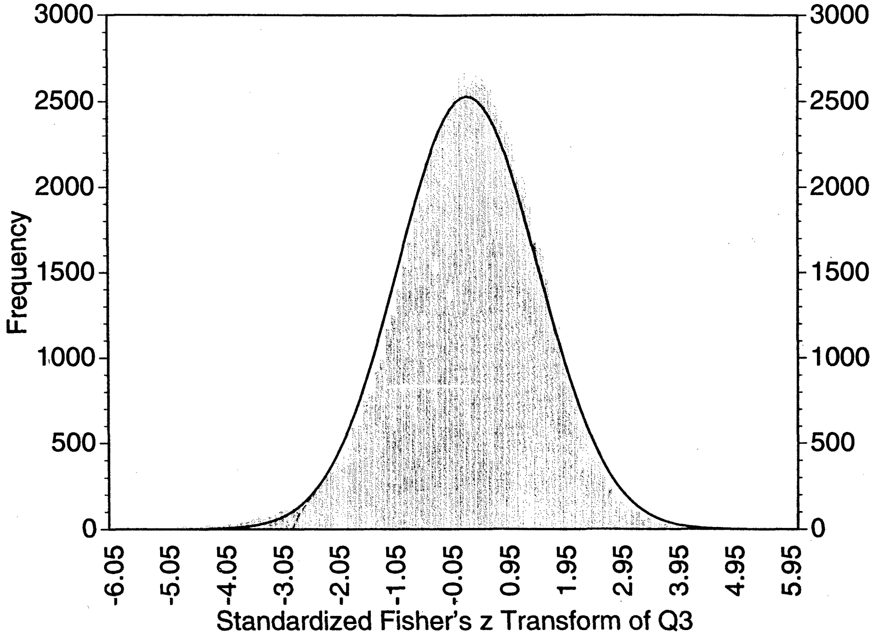

FIGURE 3. *Histogram of the empirical distribution of the standardized Fisher's z transform of* Q_3*, using the normal theory mean and variance*

Note. Q_3 is obtained from 100 replications of simulated locally independent data with 40 items and 1,000 observations. The curve is $N(0.26, 1.51)$.

correlated in Q_3 are not bivariate Gaussian. Thus, the normal theory values for the mean and variance of the sampling distribution may or may not be correct.

Figure 3 shows the sampling distribution for 78,000 values of z-transformed Q_3, standardized using those values, along with the Gaussian curve. (For Figure 3, the data involved 1,000 simulated examinees responding to 40 3PL items, with the 3PL item parameters generated randomly for each of the 100 replications). The sampling distribution is certainly bell shaped, but its approximation by the standard normal is not particularly good, especially in the (crucial) tails. The distribution is also asymmetric.

The result is that normal theory critical values do not produce empirical Type I error rates that are very close to the nominal 0.05 and 0.01 values—and they especially do not produce Type I error rates that are very close to the nominal 0.025 and 0.005 values in the two tails. Table 13 shows the empirical Type I error rates for Q_3, computed using the normal theory mean and variance for the z-transformed statistics. The values are sufficiently different from the nominal levels that we use empirically determined critical values (computed from simulation of locally independent item response data), when we consider the power of Q_3.

It should be noted that the normal theory critical values of Q_3 are probably not widely used in practice. In using Q_3 to screen items for local dependence, it is

TABLE 13

Percentages of the standardized Fisher's z transform of Q_3 outside the 0.05 and 0.01 nominal levels of a standard normal distribution for 100 replications, simulated locally independent data, for the 2PL and 3PL models, with 1,000 observations

No. of items	% < −1.96	% > 1.96	% < −2.58	% > 2.58
		2PL model		
10	2.7	42.9	1.5	25.5
40	2.9	4.1	0.9	1.0
80	2.6	3.1	0.7	0.7
		3PL model		
10	8.5	50.4	6.3	36.7
40	4.0	7.0	1.7	2.1
80	3.3	4.4	1.1	1.2

more common to use a uniform critical value of an absolute value of 0.2 for the Q_3 statistic itself (A. Fitzgerald, personal communication, April, 1994). Table 14 shows the empirical Type I error rates for that criterion, for 10-, 40-, and 80-item tests with 1,000 examinees, using the 2PL and 3PL models to generate and analyze the data. For realistic test lengths (40–80 items), the Type I error rates for a criterion of 0.2 are very small. However, the use of the 0.2 criterion would result in very low power for Q_3 in comparison to the X^2 and G^2 LD indexes.

The Power of Q_3 Relative to the G^2 LD Index

To compare the power of Q_3 with the G^2 LD index, we used critical values that were empirically obtained through the simulation of locally independent data. We noted above that the standard χ^2 critical values for the G^2 LD index are somewhat conservative, and the normal theory critical values for Q_3 are very liberal under a variety of circumstances. Using empirically determined critical values creates a level playing field for the comparison of the power of the two statistics, although it is a somewhat impractical procedure for applied use of the statistics, because it requires a rather large-scale simulation to determine the critical values for any particular sample size and test composition.

TABLE 14

Proportions of the Q_3 outside −0.2 and 0.2 for 100 replications, simulated locally independent data, for the 2PL and 3PL models, with 1,000 observations

	2PL		3PL	
No. of items	< −0.2	> 0.2	< −0.2	> 0.2
10	1.1	0.0	5.7	0.0004
40	0.0	0.0	0.0004	0.0001
80	0.000009†	0.0	0.00007††	0.00007††

Note. †3 out of 316,000. ††22 out of 316,000.

Table 15 shows the relative power of Q_3 and the G^2 LD index for data simulated with the ULD model. There are 40 items; 20 of the items load (5 each) on four specific factors, and the other 20 items load only on the general factor. For these conditions, Q_3 outperformed the G^2 LD index; at the 0.05 level, Q_3 flagged about 50% of the within-factor item pairs as locally dependent, while the G^2 LD index marked only about 35%.

Table 16 shows the relative power of Q_3 and the G^2 LD index for data simulated with the SLD model. There were 40 items, and four pairs of items are simulated locally dependent. For both statistics the proportions for the SLD pairs exceed the empirical nominal $\alpha = 0.05$ level, increasing from 30% to 99% as π_{LD} increases from 0.05 to 0.5. The power of detecting these SLD item pairs can be considered equal for the G^2 LD index and Q_3.

Why does Q_3 outperform the G^2 LD index for ULD, with the reverse being true (to a very slight extent) for SLD? Reconsider the plots of the residuals in Figures 1 and 2. An aspect of the shapes of those scatterplots that is considered by the correlational statistic (Q_3), but not by the LD indexes, is the ordered placement of the residuals along the curved lines. In underlying local dependence, the examinees have ordered values (higher or lower) on the factor defining the dependency. So, their residuals tend to fall in relatively ordered locations along those lines, higher in the (+, +) quadrant and lower in the (−, −) quadrant. Q_3 is sensitive to this ordering of the residuals and can use that information to detect the presence of the specific factor; the G^2 LD index is oblivious to that aspect of the data.

Under the SLD model for local dependence, there is no tendency for the residuals to be placed in any particular relative order, so Q_3 loses its advantage

TABLE 15

Percentages of the Fisher's z transform of Q_3 and the G^2 LD index greater than the empirical percentiles for 100 replications, simulated underlying locally dependent data, for the 3PL model

	Fisher's z from Q_3				G^2 LD index	
Category of item pair	$P_{2.5}$ (%<−2.29)	$P_{97.5}$ (%>2.50)	$P_{0.5}$ (%<−3.49)	$P_{99.5}$ (%>3.22)	P_{95} (%>3.73)	P_{99} (%>5.95)
First specific factor	0.0	55.0	0.0	35.3	38.0	23.2
Second specific factor	0.0	51.8	0.0	33.8	31.7	20.6
Third specific factor	0.1	55.6	0.0	36.6	29.1	18.6
Fourth specific factor	0.0	54.8	0.0	35.7	37.6	25.3
General factor	3.0	2.4	0.8	0.4	4.9	1.0
Cross-factors	3.0	2.3	0.5	0.4	5.1	1.1

Note. No. of items = 40, $N = 1{,}000$, weight on specific factors is $N(0.71, 0.1)$. There are 5 items on each specific factor; the remaining 20 items are loaded on the general factor only.

TABLE 16

Percentages of the Fisher's z transform of Q_3 *and the* G^2 *LD index greater than the empirical percentiles for 100 replications, simulated surface locally dependent data, for the 3PL model, 40 items,* N = 1,000

	Fisher's z from Q_3				G^2 LD index	
Category of item pair	$P_{2.5}$ (%<−2.29)	$P_{97.5}$ (%>2.50)	$P_{0.5}$ (%<−3.49)	$P_{99.5}$ (%>3.22)	P_{95} (%>3.73)	P_{99} (%>5.95)
		π_{LD} = 0.05				
LI pairs	1.9	2.1	0.3	0.3	4.7	0.9
Mixed LI & SLD pairs	1.7	2.2	0.3	0.4	4.6	1.0
Mixed SLD pairs	1.9	1.8	0.4	0.3	4.5	0.8
SLD pairs	0.3	33.3	0.0	18.3	30.5	18.8
		π_{LD} = 0.1				
LI pairs	1.9	2.1	0.3	0.4	4.5	0.6
Mixed LI & SLD pairs	1.8	2.0	0.3	0.3	4.3	0.6
Mixed SLD pairs	1.8	1.5	0.2	0.3	3.7	0.7
SLD pairs	0.0	69.0	0.0	52.3	68.0	48.0
		π_{LD} = 0.3				
LI pairs	1.7	2.4	0.3	0.5	5.2	1.0
Mixed LI & SLD pairs	2.4	1.8	0.3	0.3	4.9	1.1
Mixed SLD pairs	3.7	1.0	0.5	0.2	4.9	0.9
SLD pairs	0.0	98.3	0.0	97.3	99.0	97.5
		π_{LD} = 0.5				
LI pairs	1.7	2.3	0.3	0.5	5.5	1.0
Mixed LI & SLD pairs	3.3	1.5	0.7	0.3	5.4	1.2
Mixed SLD pairs	6.8	0.7	1.8	0.1	6.7	1.4
SLD pairs	0.0	99.3	0.0	98.5	99.8	98.8

over the G^2 LD index. At the same time, the diagnostic negative local dependence that appears in the G^2 LD index when dominating local dependence disturbs the estimation of the item parameters does not appear in the Q_3 statistics, because the relatively positive orientation of the curved lines of residuals resists any tendency of the statistic to take negative values.

Conclusion

Under the null condition of local independence, both the X^2 and G^2 LD indexes have distributions very similar to the χ^2 distribution with one degree of freedom. If we treat both of these LD indexes as χ^2 with one degree of freedom, both statistics are somewhat conservative. However, considering the issue of multiple comparisons when we test item × (item − 1)/2 of these statistics, such conservatism may be acceptable. The results also indicate that the G^2 LD index is slightly more powerful than the X^2 LD index.

Under the SLD model, these LD indexes are extremely sensitive to local dependence. Even with π_{LD} as low as 0.3, both LD indexes studied in detail detect at least 85% of the item pairs that are locally dependent. The power increases as π_{LD} increases, and approaches 1. The Type I error rate is slightly larger than the nominal 5% or 1% level when the number of items is small, and becomes the same as in the null condition when the number of items increases.

Under the ULD model, with the weakly multidimensional data we generated, the X^2 and G^2 LD indexes exhibit low sensitivity; however, both LD indexes responded to the weak multidimensionality. We learned that the weights of the items on the factors have a substantial influence on the values of both of these LD indexes. These LD indexes may not suggest the exact underlying structure (e.g., how many underlying factors), but they do suggest that the items are not unidimensional. In addition, the signs of the other two (ϕ_{diff} and τ_{diff}) LD indexes provide clues about the clustering of the items.

The LD indexes are intended to be used not for hypothesis testing but rather for diagnostic purposes. Any meaningful interpretation of the LD indexes requires skill and experience in IRT analysis and close examination of the item content. Examination of the pattern of the LD indexes across item pairs is as important as the magnitude of any single LD index.

When compared to Q_3, these LD indexes are somewhat less powerful for underlying local dependence, but equally powerful for surface local dependence. However, for Q_3 it appears that one must simulate data under the null condition to obtain optimal cutpoints. On the other hand, the G^2 LD index is nearly optimal using the 3.84 or 6.63 cutpoints of χ^2 distribution of one degree of freedom.

Future research might include more complex IRT models or models of local dependence other than the SLD and ULD models. In addition, issues other than those studied here—such as local dependence among item triplets or more than three items, and multiple category item responses—could be examined. At the present time, however, we believe that the LD indexes appear to be promising statistics for the detection of local dependence.

References

Ackerman, T. A., & Spray, J. A. (1986, April). *A general model for item dependency.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis.* Cambridge, MA: MIT Press.

Burket, G. (1993). *PARDUX Version 2.57.* Unpublished manuscript.

Chen, W. (1993). *IRT_LD: A computer program for the detection of pairwise local dependence between test items* (Research Memorandum 93-2). Chapel Hill: L. L. Thurstone Laboratory, University of North Carolina at Chapel Hill.

Chen, W. (1994). *Local dependence indices for item pairs using item response theory.* Unpublished master's thesis, University of North Carolina at Chapel Hill.

Haberman, S. J. (1978). *Analysis of qualitative data: Vol. I. Introductory topics.* New York: Academic Press.

Kingston, N. M., & Dorans, N. J. (1982). *The feasibility of using item response theory as a psychometric model for the GRE aptitude test* (GRE Board Professional Report GREB 79-12P, ETS Research Report 82-12). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., & Bock, R. D. (1989). *BILOG 3: Item analysis and test scoring with binary logistic models.* Mooresville, IN: Scientific Software, Inc.

Mood, A. M., Graybill, F. A., & Boes, D. C. (1985). *Introduction to the theory of statistics* (3rd ed.). New York: McGraw-Hill.

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

Thissen, D., Bender, R., Chen, W., Hayashi, K., & Wiesen, C. A. (1992). *Item response theory and local dependence: A preliminary report* (Research Memorandum 92-2). Chapel Hill: L. L. Thurstone Laboratory, University of North Carolina at Chapel Hill.

Thissen, D., & Chen, W. (1993). *Item response theory and local dependency: An interim report* (Research Memorandum 93-3). Chapel Hill: L. L. Thurstone Laboratory, University of North Carolina at Chapel Hill.

van den Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. *Psychometrika, 47,* 123–140.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5,* 245–262.

Yen, W. M. (1984). Effect 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.

Authors

WEN-HUNG CHEN is Research Associate I, ACT, 2201 N. Dodge Street, Iowa City, IA 52243-0168; chenl@act.org. He specializes in item response theory.

DAVID THISSEN is Professor, Department of Psychology, University of North Carolina, CB#3270, Davie Hall, Chapel Hill, NC 27599-3270; dthissen@email.unc.edu. He specializes in quantitative psychology and measurement.

Received August 16, 1994
Revision received July 6, 1995
Second revision received January 18, 1996
Accepted February 13, 1996