Effects of Local Item Dependence on the Fit and Equating Performance of the Three-Parameter Logistic Model

Wendy M. Yen
CTB/McGraw-Hill

Unidimensional item response theory (IRT) has become widely used in the analysis and equating of educational achievement tests. If an IRT model is true, item responses must be locally independent when the trait is held constant. This paper presents several measures of local dependence that are used in conjunction with the three-parameter logistic model in the analysis of unidimensional and two-dimensional simulated data and in the analysis of three mathematics achievement tests at Grades 3 and 6. The measures of local dependence (called Q_2 and Q_3) were useful for identifying subsets of items that were influenced by the same factors (simulated data) or that had similar content (real data). Item pairs with high Q_2 or Q_3 values tended to have similar item parameters, but most items with similar item parameters did not have high Q_2 or Q_3 values. Sets of locally dependent items tended to be difficult and discriminating if the items involved an accumulation of the skills involved in the easier items in the rest of the test. Locally dependent items that were independent of the other items in the test did not have unusually high or low difficulties or discriminations. Substantial unsystematic errors of equating were found from the equating of tests involving collections of different dimensions, but substantial systematic errors of equating were only found when the two tests measured quite different dimensions that were presumably taught sequentially.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 8, No. 2, Spring 1984, pp. 125–145

0146-6216/84/020125-21$2.30

The fundamental assumption of the three-parameter logistic model is that

$$P_i(\theta_k) = c_i + \frac{1 - c_i}{1 + \exp[-1.7a_i(\theta_k - b_i)]}, \quad (1)$$

where $P_i(\theta_k)$ is the probability that the kth examinee passes the ith item,
θ_k is the trait value for the kth examinee, and
a_i, b_i, and c_i are the discrimination, difficulty, and lower asymptote for the ith item.

Local item independence follows from Equation 1; if the model is true, the probability of correctly answering both items i and j is

$$P_{ij}(\theta_k) = P_i(\theta_k)P_j(\theta_k) \quad . \quad (2)$$

Equation 2 defines pairwise local independence. If the model is true, then all higher order extensions of Equation 2 will hold as well.

In examining the fit of the model it is common to compare model predictions (Equation 1) with observations. This comparison traditionally has been between observed and predicted item characteristic curves, and there has been little direct investigation of local independence. There are many factors that can affect the fit of a model. One of the most obvious is multidimensionality. For many tests there is an a priori suspicion that test performance is not unidimensional and that a unidimensional model is not appropriate. This concern is particularly strong

for achievement tests where several different types of content are tested (Traub, 1983). Such multidimensionality would be expected to have a particularly strong effect on the local independence of at least some of the items. Given this concern, it is desired to measure local dependence and examine its impact on the use of a unidimensional model—the three-parameter logistic model.

In this paper several measures of local dependence are proposed, and they are applied to simulated data and to data from six mathematics achievement tests. The effects of local dependence on a traditional measure of model fit are examined. Local dependence is also related to item parameter values and to the quality of test equatings.

Method

Fit Measures

Q_1. Traditional measures of fit in item response theory (IRT) typically involve comparisons between observed and predicted item characteristic curves using a chi-square test such as Q_1 (Yen, 1981):

$$Q_{1i} = \sum_{r=1}^{10} \frac{N_r(O_{ir} - E_{ir})^2}{E_{ir}(1 - E_{ir})} \quad . \tag{3}$$

Q_{1i} is the fit of the ith item, and it is distributed approximately as a chi-square variable with seven degrees of freedom when the three-parameter model is true. To obtain Q_{1i} examinees are rank ordered on the basis of their trait estimates and then divided into deciles (10 cells with approximately equal numbers of examinees per cell). N_r is the number of examinees in cell r, O_{ir} is the observed proportion of examinees in cell r that correctly answers item i, and E_{ir} is the predicted proportion of examinees in cell r that correctly answers item i. The predictions are obtained by using trait and item parameter estimates in Equation 1 and summing over examinees in cell r:

$$E_{ir} = \frac{1}{N_r} \sum_{k \in r}^{N_r} \hat{P}_i(\hat{\theta}_k) \quad . \tag{4}$$

The fit of the entire test is $Q_1 = \Sigma_{i=1}^{I} Q_{1i}$, with I items in the test.

Q_2. Van den Wollenberg (1982) proposed a fit measure for the Rasch model, called Q_2, that he found is sensitive to multidimensionality. To obtain Q_2, a contingency table is created for each pair of items for each (overall test) score group (cell) as in Table 1. In this table N_r is the number of examinees in cell r, N_{ijr} is the number of examinees in the cell that correctly answered both items i and j, N_{ir} is the number of examinees that correctly answered item i, $N_{i\bar{j}r}$ is the number of examinees that correctly answered item i and incorrectly answered item j, and so on.

Van den Wollenberg's fit statistic for the rth cell is

$$\begin{aligned} Q_{2ijr} = {} & \frac{[N_{\bar{i}\bar{j}r} - \mathrm{E}(N_{\bar{i}\bar{j}r})]^2}{\mathrm{E}(N_{\bar{i}\bar{j}r})} \\ & + \frac{[N_{\bar{i}jr} - \mathrm{E}(N_{\bar{i}jr})]^2}{\mathrm{E}(N_{\bar{i}jr})} \\ & + \frac{[N_{i\bar{j}r} - \mathrm{E}(N_{i\bar{j}r})]^2}{\mathrm{E}(N_{i\bar{j}r})} \\ & + \frac{[N_{ijr} - \mathrm{E}(N_{ijr})]^2}{\mathrm{E}(N_{ijr})} \quad . \end{aligned} \tag{5}$$

The overall fit measure for item pair i,j is the sum of the Q_{2ijr} values over the R cells. Van den Wollenberg obtained the expected values for Q_2 using a procedure specifically associated with the Rasch model.

In generalizing Q_2 for the three-parameter model, 10 cells were used as defined for Q_1. The expectations were defined as

$$\mathrm{E}(N_{ijr}) = N_{ir}N_{jr}/N_r \quad ,$$

$$\mathrm{E}(N_{\bar{i}jr}) = N_{\bar{i}r}N_{jr}/N_r \quad ,$$

$$\mathrm{E}(N_{i\bar{j}r}) = N_{ir}N_{\bar{j}r}/N_r \quad ,$$

and

$$\mathrm{E}(N_{\bar{i}\bar{j}r}) = N_{\bar{i}r}N_{\bar{j}r}/N_r \quad . \tag{6}$$

This characterization of Q_2 appears consistent with Lord's (1953, p. 545) suggested procedure for examining local independence.

Q_2 takes the form of a Pearson chi-square statistic (Hays, 1973). There are 10 independent cells (observations) with one degree of freedom per cell, leading to the conclusion that Q_2 should follow a chi-square distribution with 10 degrees of freedom

Table 1
Number of Examinees in Cell r With Each Pattern of Item Scores

Item i Score	Item j Score 0	1	Total
0	$N_{\bar{i}\bar{j}r}$	$N_{\bar{i}jr}$	$N_{\bar{i}r}$
1	$N_{i\bar{j}r}$	N_{ijr}	N_{ir}
Total	$N_{\bar{j}r}$	N_{jr}	N_{r}

when the null hypothesis (local independence) is true. The determination of the cell boundaries is dependent on the observed distribution of estimated trait values, which may affect the distribution of Q_2 and its degrees of freedom. The null distribution of Q_2 will be examined with simulation data.

Using the expectations in Equation 6, $Q_{2ijr} = N_r\phi^2_{ijr}$, where ϕ_{ijr} is the phi coefficient (correlation) between item scores in cell r. Binary scores (such as those for items) are pairwise independent if and only if they are uncorrelated, because the covariance of items i and j is $P_{ij} - P_iP_j$. Pairwise independence does not preclude higher order dependencies, though in most testing applications it is difficult to imagine important higher-order dependencies that are not reflected in pairwise dependence.

Although Q_2 offers a method of determining where local dependence exists, it does not reflect the positive or negative direction of the local dependence. It is easy to imagine positive local dependence when two or more items measure special traits that do not appear in the remainder of the test. Negative local dependence can appear between two sets of items that measure different traits. For example, imagine that the trait estimated by the unidimensional model is the average of two underlying traits and that to look for local dependence examinees are sorted into cells on the basis of this average. An examinee can be placed in a cell for moderate scorers by having moderate scores on both underlying traits, by having a low score on Trait 1 and a high score on Trait 2, by having a high score on Trait 1 and a low score on Trait 2, and so forth. This procedure will reveal a negative relationship between performance on items measuring Trait 1 and items measuring Trait 2.

To estimate the direction of the relationship between items i and j, $\Sigma_{r=1}^{10}\phi_{ijr}$ was obtained. The sign of this sum was then applied to Q_2 to create "signed Q_2."

Q_3. The disadvantage of Q_2 and signed Q_2 is that the number and definition of the cells are arbitrary. Kingston and Dorans (1982) examined local dependence using a statistic that does not require cells—the correlation of item scores with the trait estimate partialled out ($r_{ij\cdot\hat{\theta}}$). The disadvantage of this statistic is that it removes only the linear relationship between item scores and traits when it is known that items and traits have a nonlinear, logistic relationship. An alternative to $r_{ij\cdot\hat{\theta}}$ was examined that removed the nonlinear effects of $\hat{\theta}$ from the item scores. Define

$$d_{ik} = u_{ik} - \hat{P}_i(\hat{\theta}_k) \quad , \tag{7}$$

where u_{ik} is the score of the kth examinee on the ith item. Then, the correlation (taken over examinees) of these scores is

$$Q_{3ij} = r_{d_id_j} \quad . \tag{8}$$

Because d_{ik} and d_{jk} are random error scores when the three-parameter model is true, they may be distributed approximately as bivariate normal variables with a zero correlation. Thus, when the model is true, Fisher's r-to-z transformation of Q_3 may be distributed as a normal variable with mean equal to zero and variance equal to $1/(N - 3)$, where there are N examinees involved in the calculation of the correlation. Kingston and Dorans (1982) noted that because item scores are involved in the calculation of $\hat{\theta}$, $r_{ij\cdot\hat{\theta}}$ values will tend to be slightly negative due to part-whole contamination. A similar effect may occur with Q_3. The distribution of the z transformation of Q_3 will be examined with simulated data.

Parameter Estimation

LOGIST 5 Version 1.0 (Wingersky, Barton, & Lord, 1982) was used to estimate item parameters and trait values for the three-parameter logistic

model. The default options were used in running LOGIST. In estimating item parameters, omits were treated as incorrect answers and not-reached items were treated as "not reaches."

Q_2, signed Q_2, and Q_3 were obtained for each pair of items in each LOGIST run. Also, the average of the Q_1 values for each pair of items was obtained and called $\overline{Q}_1$. Examinees who did not reach an item were not included in the calculation of these statistics for that item.

Simulated Data

Tests and examinees. To examine the null distribution of Q_2 and the z transformation of Q_3, simulated data were used. A simulated unidimensional test was created by using the parameters of the first 20 items of the Level F Reading Vocabulary test from the *Comprehensive Tests of Basic Skills, Form U* (CTBS/U; CTB/McGraw-Hill, 1981). These parameters were obtained from that test's standardization (CTB/McGraw-Hill, 1982a; Yen, 1983b). An additional unidimensional simulated test was created using the parameters of the first 40 items of that Reading Vocabulary test. Item responses for these two tests were generated from the three-parameter model (Equation 1) for 1000 simulees as described by Yen (in press).

Simulated data for three 30-item multidimensional tests were also examined. Item responses for these tests were generated using a two-dimensional three-parameter logistic model whose general form is described by Reckase and McKinley (1983):

$$P_i(\theta_{k1},\theta_{k2}) = c_i + \frac{1 - c_i}{1 + \exp[-1.7 \sum_{s=1}^{2} a_{is}(\theta_{ks} - b_{is})]} . \quad (9)$$

The parameters and traits in this model are the same as in Equation 1 with the addition of the subscript s, which specifies the two dimensions of the model.

The present study used item parameters from the "Easy" tests in "Configurations 1, 2, and 3" for the "$\overline{b}_2 - \overline{b}_1 = 0$" condition described by Doody-Bogan and Yen (1983). Item responses generated for 2,000 "Low and Medium Ability" simulees in that study were used in the present study.

Table 2 displays the a_{is} values for the three multidimensional simulated conditions. These values indicate how the items were differentially affected by the two dimensions. In Configuration 1, the correlation between θ_1 and θ_2 was approximately .60, and the a_{is} and b_{is} values were chosen to have essentially zero correlations within and across dimensions. In Configuration 2, the correlation between θ_1 and θ_2 was approximately .50, and the b_{is} values were chosen to be essentially uncorrelated with each other and with the a_{is} values. In Configuration 3, the correlation between θ_1 and θ_2 was approximately .50, and the b_{is} values had a slightly negative correlation with the a_{is} values for the first dimension ($r = -.31$) and a slightly positive correlation ($r = .09$) for the second dimension. For every dimension and configuration, $\overline{b} = -.03$ and $S_b = .80$.

In running LOGIST for the simulated data the number of answer choices was set at four.

Real Data

Tests and examinees. The CTBS/U Mathematics Computation (MC) and Mathematics Concepts and Applications (MC&A) tests were used. Level E (Grade 3) and Level G (Grade 6) were employed. The third- and sixth-grade students were tested as part of the Fall 1980 national standardization of CTBS/U. In addition to the CTBS/U tests, examinees took the *Diagnostic Mathematics Inventory* (DMI; Gessel, 1975a). DMI Level B was given at Grade 3 and Level E was given at Grade 6. The DMI items were classified as MC if they involved only adding, subtracting, multiplying, or dividing. All other DMI items were classified as MC&A items. Table 3 contains the numbers of items and examinees.

The CTBS/U MC items have five answer choices and the CTBS/U MC&A items have four answer choices. The DMI Level B items have five answer choices (except for one item with six answer choices), and the DMI Level E items have 8 (66% of the items), 9 (11%), or 10 (23%) answer choices.

For each grade, the CTBS/U and DMI MC items were analyzed together by LOGIST, and the CTBS/U and DMI MC&A items were analyzed together.

Table 2

Item Discriminations for Multidimensional Simulations

Item (i)	Configuration 1		Configuration 2		Configuration 3	
	a_{i1}	a_{i2}	a_{i1}	a_{i2}	a_{i1}	a_{i2}
1	.8	.6	.6	1.8	.5	.6
2	1.0	.7	.8	1.1	.6	.5
3	1.1	.6	.9	.9	.6	.6
4	1.2	.5	1.1	.5	.7	.7
5	1.4	.9	.7	.0	.7	.7
6	1.6	.9	.8	.0	.8	.0
7	1.8	1.1	.8	.0	.8	.0
8	2.0	.8	.9	.0	.8	.0
9	.5	.9	.9	.0	.8	.0
10	.8	1.0	1.0	.0	.9	.0
11	.8	1.1	1.0	.0	.9	.0
12	.9	1.1	1.1	.0	.9	.0
13	.7	1.2	1.1	.0	.9	.0
14	1.0	1.3	1.2	.0	1.0	.0
15	.6	1.4	1.2	.0	1.0	.0
16	1.3	1.6	1.6	.0	1.0	.0
17	1.2	1.8	2.0	.0	1.0	.0
18	1.0	2.0	.0	.6	1.0	.0
19	.6	.7	.0	.7	1.1	.0
20	.7	.8	.0	.8	1.1	.0
21	.8	.8	.0	.8	1.1	.0
22	.9	.8	.0	.9	1.1	.0
23	.9	.9	.0	1.0	1.2	.0
24	.9	1.0	.0	1.0	1.2	.0
25	1.0	1.0	.0	1.0	1.2	.0
26	1.0	1.1	.0	1.1	1.3	.0
27	1.1	1.0	.0	1.2	1.4	.0
28	1.1	1.0	.0	1.2	1.6	.0
29	1.1	1.2	.0	1.3	1.8	.0
30	1.2	1.2	.0	1.4	2.0	.0

Note. For explanatory purposes the item order has been changed from that in Doody-Bogan and Yen (1983).

In addition, the MC and MC&A items were analyzed together. The number of answer choices for each test was set at the value described above; the number of answer choices was set at five for DMI Level B and at eight for DMI Level E.

Item configurations. Values of fit statistics were compared for groups of items using the Kruskal-Wallis test (Siegel, 1956, pp. 184–185). Items were grouped into sets that appeared most likely to show local dependence, and the null hypothesis of the Kruskal-Wallis test was that the fit statistics within these sets arose from the same distribution as the fit statistics calculated across item pairs from different sets. The Kruskal-Wallis test comparing within-set and between-set fit statistics has a chi-square distribution with 1 degree of freedom under the null hypothesis.

Table 4 contains the definitions of the item configurations. Splitting the items by test (CTBS/U vs. DMI) is the first configuration for every grade-by-content category. The *CTBS, Forms U and V, Test Coordinator's Handbook* (CTB/McGraw-Hill,

Table 3
Numbers of Items and Examinees

	CTBS/U		DMI		
Grade	MC	MC&A	MC	MC&A	N
3	36	40	14	40	2184
6	40	45	48	61	2198

1982b) and the *DMI Teacher's Guide* (Gessel, 1975b) were referred to for classifications of items into category objectives. For example, the Grade 3 MC test for CTBS/U has the following category objectives:

Add whole numbers,
Add decimals or fractions,
Subtract whole numbers,
Subtract decimals or fractions,
Multiply whole numbers,
Divide whole numbers.

The Grade 3 MC test for DMI has the following category objectives:

Addition of whole numbers without regrouping (i.e., carrying),
Addition of whole numbers with regrouping,
Subtraction of whole numbers without regrouping (i.e., borrowing),
Subtraction of whole numbers with regrouping,
Multiplication of whole numbers.

As can be seen in Table 4, various combinations and modifications of these category objectives were used to group items that appeared most likely to be locally dependent. For the MC&A tests reconciliation of the category objectives for CTBS/U and DMI was more difficult than for the MC tests because different objective structures and terminology are used for these two tests, but sets of items most likely to be locally dependent were identified. Note that the same names were sometimes used for different item sets in different configurations (e.g., Sequences for Grade 3 MC&A Data Sets 5 and 6).

In Data Sets 6 and 15 items were grouped that showed high Q_2 values. For example, in Data Set 6 the Graph Reading items were two CTBS/U items linked to the same graph, the Clock items all involved reading clocks, the Nearest Ten items involved rounding to the nearest 10, and so forth. The Other items were all the remaining items that did not show high Q_2 values.

Trait estimation and comparison. Trait level estimates were obtained for different subsets of items using the item parameter estimates described above. In estimating trait levels, omits and not reaches were treated as wrong answers. Floors and ceilings for the trait estimates were obtained as described by Yen (1983b).

Grade 3 MC&A (for both CTBS/U and DMI) and Grade 6 MC (for both CTBS/U and DMI) were selected for the examination of trait estimates. For Grade 3 MC&A, trait estimates based on CTBS/U items were compared with trait estimates based on DMI items. Three additional trait estimates were obtained. The first was composed of the 27 locally dependent items shown in Data Set 6 in Table 4. The second trait estimate (Control 1) was composed of 26 items not included in the first trait estimate, and the third trait estimate (Control 2) was based on 25 items not included in the first or second trait estimate. Because the items in the locally dependent set tended to be adjacent pairs, the items chosen for Controls 1 and 2 also tended to be adjacent pairs.

For Grade 6 MC, trait estimates based on CTBS/U items were compared with trait estimates based on DMI items. Three traits based on locally dependent items were also examined: items involving computations with whole numbers (36 items), items involving decimals (36 items), and items involving fractions (16 items). The computations involving different types of numbers tended to be interspersed

Table 4

Item Configurations for Examining Local Dependence

Data Set	Grade	Test		Item Sets
		CTBS/U	DMI	
1	3	MC	MC	CTBS/U(36);DMI(14)
2	3	MC	MC	Add(16);Subtract(16); Mult(12);Divide(6)
3	3	MC	MC	Add without regroup(7);Add with regroup(9);Subtract without regroup(9);Subtract with regroup(7);Mult by 5 (4);Other mult(8);Divide ÷ (3);Divide $\sqrt{\ }$ (3)
4	3	MC&A	MC&A	CTBS/U(40);DMI(40)
5	3	MC&A	MC&A	Counting & Matching(6);Numeration(10);Number theory(10); Measurement(17);Geometry(7); Problem solving(10);Number Sentences(17);Sequences(3)
6	3	MC&A	MC&A	Graph reading(2);Clock(3); Nearest ten(2);Place value (2);Number sentences(2);Even numbers(2);Before(2);Number line(2);Multiply(4);Fractions (2);Sequences(2);Shapes(2); Other(53)
7	3	MC	MC&A	CTBS/U(36);DMI(40)
8	3	MC&A	MC	CTBS/U(40);DMI(14)
9	6	MC	MC	CTBS/U(40);DMI(48)
10	6	MC	MC	Add(21);Subtract(22);Mult (23);Divide(22)
11	6	MC	MC	Whole numbers(36);Decimals (36);Fractions(16)
12	6	MC	MC	Add whole(3);Subtract whole (6);Mult whole(11);Divide whole(16);Add dec(12);Subtract dec(10);Mult dec(9); Divide dec(5);Add fract(6); Subtract fract(6);Mult & Divide fract(4)
13	6	MC&A	MC&A	CTBS/U(45);DMI(61)
14	6	MC&A	MC&A	Number sentences(18);Measurement(18);Geometry(18);Problem solving(19);Numeration(18); Number theory(10);Sequences (2);Inequalities,Odds,Multiples(3)
15	6	MC&A	MC&A	Greater than(2);Fractions(8); Estimation(4);Sequences, Inequalities,Odds(4);Conventional measurement(2);Place value DMI(2);Expanded notation(2);Place value CTBS(2); Other(80)
16	6	MC	MC&A	CTBS/U(40);DMI(61)
17	6	MC&A	MC	CTBS/U(45);DMI(48)

Note. The number of items in a set is in parentheses following the name of the set.

in CTBS/U and grouped in DMI. Three additional traits were: Control 1, a set of 36 items interspersed across both CTBS/U and DMI; Control 2, a set of 36 items not included in Control 1; and Control 3, a 16-item subset of the items in Control 2.

Several statistics were used to compare different trait estimates: (1) the correlation, r; (2) standardized difference between means; (3) ratio of standard deviations; (4) standardized root mean squared difference (SRMSD); and (5) an estimate of local bias (Yen, 1983a). The standardized difference between means is the difference in mean scores for the two sets of traits divided by a pooled estimate of the standard deviation: $\bar{S} = [(S_1^2 + S_2^2)/2]^{1/2}$, where S_1^2 and S_2^2 are the variances of the two sets of scores. The standardized root mean squared difference is the square root of the mean squared difference between examinees' trait estimates, divided by $\bar{S}$.

To estimate local bias, examinees were rank ordered on the basis of their scores and grouped into quintiles (five cells with equal numbers of examinees in the cells). The average of the two scores being compared was used for the grouping. Within each cell the standardized difference between means was found. The average over cells of the absolute values of the standardized mean differences was taken as the summary measure of local bias.

Results

Simulations

Unidimensional data. The mean and variance of Q_1 for the unidimensional simulation were 10.8 and 58.1 for the 20-item test and 7.1 and 12.5 for the 40-item test. A goodness-of-fit test was conducted to determine if the Q_1 values had chi-square distributions with 7 degrees of freedom. In this test 10 cells were created with equal numbers of Q_1 values expected in each cell. The expectations were generated using a chi-square distribution with 7 degrees of freedom. The significance probabilities for these tests were .07 for the 20-item simulation and .99 for the 40-item simulation.

For the unidimensional simulation the mean and variance of the Q_2 values were 10.3 and 19.4 for the 20-item test and 9.4 and 23.9 for the 40-item test. A goodness-of-fit test was conducted to determine if the Q_2 values had chi-square distributions with 10 degrees of freedom. In this test 10 cells were created with equal numbers of Q_2 values expected in each cell. The expectations were generated using a chi-square distribution with 10 degrees of freedom. The significance probabilities for these tests were .36 for the 20-item simulation and $<.001$ for the 40-item simulation. The Q_2 values for the 20-item simulation looked like they had a chi-square distribution with 10 degrees of freedom, but the Q_2 values for the 40-item simulation had more small values of Q_2 than expected for a chi-square distribution with 10 degrees of freedom. Thus, though the null hypothesis about the distribution of the Q_2 value was upheld only for the 20-item test, the deviations from the expected distribution for the 40-item test were not in the upper tail of the distribution where hypothesis testing is done. Large values of Q_2 that would be unlikely for a chi-square distribution with 10 degrees of freedom were unlikely in the simulated data.

The means of the Q_3 statistics were $-.05$ for the 20-item test and $-.02$ for the 40-item test. The significance probabilities of the chi-square tests that the z transformations of these Q_3 values came from normal distributions with a mean of zero and a variance of $1/(N-3)$ were $<.001$ for both the 20- and 40-item tests. The correlations were more negative than expected by the null hypothesis.

Two-dimensional data. The means and variances of the Q_1 values for the two-dimensional simulations were 13.4 and 121.7 for Configuration 1, 9.7 and 20.0 for Configuration 2, and 8.0 and 15.8 for Configuration 3. The significance probabilities of the tests that these Q_1 values came from a chi-square distribution with 7 degrees of freedom were $<.001$ for Configuration 1, .01 for Configuration 2, and .86 for Configuration 3. For Configurations 1 and 2 there were more large Q_1 values than expected from the null distribution. The means and variances of the Q_2 values for the two-dimensional simulated data were 10.1 and 34.6 for Configuration 1, 17.4 and 75.7 for Configuration 2, and 11.0 and 29.2 for Configuration 3.

The significance probabilities of the tests that these Q_2 values came from a chi-square distribution with 10 degrees of freedom were .10 for Configuration 1, <.001 for Configuration 2, and .01 for Configuration 3. For Configurations 2 and 3 there were more large Q_2 values than expected from the null distribution. The mean Q_3 value was $-.03$ for each of the three configurations. The significance probabilities that the z transformations of the Q_3 values came from normal distributions with zero means and variances of $1/(N-3)$ were <.001 for each of the three configurations. For Configurations 1 and 3 there were more large negative z values than expected, and for Configuration 2 there were more extreme negative and positive z values than expected from the null distribution.

For Configurations 1, 2, and 3 the correlations between $\overline{Q}_1$ and Q_2 were $-.11$, $-.09$, and .00, the correlations between $\overline{Q}_1$ and Q_3 were $-.05$, $-.03$, and $-.10$, and the correlations between signed Q_2 and Q_3 were .70, .94, and .75. While $\overline{Q}_1$ appeared to measure a different aspect of item fit than Q_2 and Q_3, these latter two statistics appeared to measure very similar properties of the items.

Items were grouped into sets that might be locally dependent, and the Kruskal-Wallis test was performed to see if the fit values within sets came from the same distribution as the fit values between sets. In Configuration 1 items 1–8 formed Set 1, items 9–18 formed Set 2, and items 19–30 formed Set 3. In Configuration 2 items 1–4 formed Set 1, items 5–17 formed Set 2, and items 18–30 formed Set 3. In Configuration 3 items 1–5 formed Set 1 and items 6–30 formed Set 2. Table 5 presents the results of these tests. The Between means were based on all item pairs across item sets.

The pattern of results for $\overline{Q}_1$ does not show significantly higher values within item sets than between sets, and the results for $\overline{Q}_1$ do not follow the pattern of results for the other fit statistics. Comparing the results for Q_2 and signed Q_2 for Configuration 1, it is apparent that there were both positive and negative local dependence within the item sets; the level of Q_2 and Q_3 values were within the range found for the unidimensional simulations.

Table 5

Mean $\overline{Q}_1$, Q_2, Signed Q_2, and Q_3 Values and Kruskal-Wallis (K-W) Statistics for Item Configurations for Two-Dimensional Simulated Data

Configuration	Item Set	Number of Items	$\overline{Q}_1$	Q_2	Signed Q_2	Q_3
1	1	8	11	10	-5	-.02
	2	10	12	9	-3	-.02
	3	12	16	10	-7	-.03
	Between		13	10	-6	-.03
	K-W		<1	2	<1	<1
2	1	4	11	24	-24	-.08
	2	13	11	17	15	.04
	3	13	8	15	3	.00
	Between		10	19	-15	-.06
	K-W		<1	12	139	155
3	1	5	6	13	13	.03
	2	25	8	11	-7	-.03
	Between		7	11	-8	-.03
	K-W		7	<1	4	3

In Configuration 2 all the local dependence within Item Set 1 was negative, and almost all the local dependence within Item Set 2 was positive; Item Set 3 contained mixed positive and negative local dependence. The within-set and between-set levels of local dependence differed significantly ($p < .001$) for Q_2, signed Q_2, and Q_3.

In Configuration 3 all the local dependence within Item Set 1 was positive, and almost all the local dependence within Item Set 2 and between sets was negative. The significance probabilities for the Kruskal-Wallis tests were .04 for the signed Q_2 values and .08 for the Q_3 values.

Real Data

Fit measures. The correlations of $\overline{Q}_1$ with Q_2 ranged from .01 to .02, and the correlations of $\overline{Q}_1$ with Q_3 ranged from $-.05$ to .03. There was essentially no relationship between $\overline{Q}_1$ and Q_2 or Q_3. The correlation between signed Q_2 and Q_3 ranged from .78 to .93, and these two statistics appeared to measure very similar properties of the item pairs.

Significance tests were conducted to determine if the fit statistics followed their null distributions. The probability that the Q_1 values came from a chi-square distribution with 7 degrees of freedom was $<.005$ for every real data set. The probability that the Q_2 values came from a chi-square distribution with 10 degrees of freedom was $<.001$ for every real data set, and the probability that the z transformations of the Q_3 statistics came from a normal distribution with mean zero and variance $1/(N-3)$ was $<.001$ for every real data set. There were more large Q_1 and Q_2 values than expected by their null distributions, and the z statistics tended to be more extreme than expected.

Tables 6 and 7 contain mean values of $\overline{Q}_1$, Q_2, signed Q_2, and Q_3 for the Data Sets in Grade 3 (Table 6) and Grade 6 (Table 7). The Between means were based on all item pairs across item sets. Also in Tables 6 and 7 are the Kruskal-Wallis statistics testing that the within-set fit values arose from the same population as the between-set fit values. Kruskal-Wallis values greater than 6.6 are significant at the .01 level and values greater than 10.8 are significant at the .001 level.

It is clear from Tables 6 and 7 that $\overline{Q}_1$ had very different patterns of results from the other fit measures. Item sets that had high $\overline{Q}_1$ values did not necessarily have high Q_2 or Q_3 values. In other words, item pairs that appeared locally dependent using Q_2, signed Q_2, or Q_3 did not individually tend to have poor fit between model predictions and observations of item characteristic curves.

The signed and unsigned Q_2 values can be compared to reveal how much local dependence there is and its direction. The change in magnitude from the unsigned to the signed Q_2 reflects how much of the local dependence is negative. The signed Q_2 values and the Q_3 values are more useful than unsigned Q_2 values for comparing the within-set and between-set local dependence, because the between-set unsigned Q_2 values can be large due to negative local dependence. Almost every within-set mean signed Q_2 value and mean Q_3 value was positive and every between-set mean signed Q_2 value and mean Q_3 value was negative, indicating that the item sets did tend to define different traits. Every signed Q_2 Kruskal-Wallis test and every Q_3 Kruskal-Wallis test was significant at less than the .01 level except for Data Sets 1 and 6.

For Data Set 2, which is based on Grade 3, Dividing items were highly locally dependent. However, by Grade 6 the same type of item grouping (Data Set 10) did not show a high degree of local dependence for the Dividing items. Data Set 3 showed a great deal of local dependence within item sets such as Adding with regrouping, Subtracting with regrouping, Multiplying by 5, Dividing with $\div$ notation, and Dividing with $\sqrt{\ }$ notation. For Data Set 11 Fractions computations were locally dependent, and Data Set 12 showed that Adding decimals, Adding fractions, and Subtracting fractions had particularly high local dependence.

For Data Sets 5 and 14 there was not an extremely high degree of local dependence in the a priori category objectives except for Sequences (Grades 3 and 6) and Number theory (Grade 6 only). However, by examining the Q_2 values and regrouping the items into Data Sets 6 and 15, some very highly locally dependent item sets were iden-

Table 6

Mean $\bar{Q}_1$, Q_2, Signed Q_2, and Q_3 Values and Kruskal-Wallis (K-W) Statistics for Data Sets for Grade 3

Data Set & Item Type	$\bar{Q}_1$	Q_2	Signed Q_2	Q_3	Data Set & Item Type	$\bar{Q}_1$	Q_2	Signed Q_2	Q_3
Data Set 1					Data Set 6				
CTBS/U	28	27	1	-.01	Graph read	10	124	124	.26
DMI	16	30	14	.02	Clock	34	78	78	.19
Between	22	16	-7	-.03	Nearest ten	17	54	54	.09
K-W	53	23	<1	2	Place value	20	208	208	.28
Data Set 2					Number sent	14	60	60	.13
Add	25	33	24	.04	Even numbers	25	146	146	.22
Subtract	22	29	17	.03	Before	14	240	240	.36
Multiply	31	35	32	.08	Number line	23	91	91	.18
Divide	18	175	175	.29	Multiply	19	71	71	.15
Between	25	17	-12	-.04	Fractions	14	639	639	.55
K-W	<1	23	285	309	Sequences	6	155	155	.25
Data Set 3					Shapes	13	76	76	.17
Add w/o rg	23	21	20	.03	Other	19	11	-2	-.01
Add with rg	26	78	73	.14	Between	19	11	-3	-.01
Sub w/o rg	28	26	23	.05	K-W	17	<1	8	6
Sub with rg	15	76	76	.15	Data Set 7				
Mult by 5	25	50	50	.13	CTBS/U	24	26	11	.01
Other mult	34	28	24	.07	DMI	17	13	4	.00
Divide ÷	27	207	207	.32	Between	20	12	-7	-.03
Divide $\sqrt{\ }$	8	206	205	.31	K-W	1	3	227	334
Between	24	19	-9	-.03	Data Set 8				
K-W	2	71	249	249	CTBS/U	20	12	-2	-.01
Data Set 4					DMI	20	43	39	.07
CTBS/U	19	11	0	.00	Between	20	12	-7	-.03
DMI	19	12	5	.01	K-W	<1	<1	75	118
Between	19	11	-5	-.02					
K-W	<1	<1	250	460					
Data Set 5									
Count & match	12	23	21	.03					
Numeration	19	16	4	.00					
Number theory	23	14	5	.00					
Measurement	22	12	1	.00					
Geometry	18	19	16	.04					
Problem solv	18	13	4	.00					
Number sent	19	19	11	.01					
Sequences	9	59	59	.10					
Between	19	11	-3	-.01					
K-W	6	<1	62	59					

tified. Although all the item pairs with similar content were not necessarily locally dependent, all the item pairs with high Q_2 or Q_3 values had clear content similarities. It should be noted that the Kruskal-Wallis tests are not helpful with Data Sets 6 and 15 because of the Other group. The Other group, while classified as an item set, was actually a collection of items that did not have noticeably high Q_2 values. Thus, when the Kruskal-Wallis test was performed, the Other Q_2 or Q_3 values were very similar to the Between Q_2 or Q_3 values. Also, because the Q_2 values were used in the classification of items into sets for Data Sets 6 and 15, the Q_2 values had to be higher for those sets. The mean Q_2 and Q_3 values for Data Sets 6 and 15 are of interest, however, because they show the amount

Table 7

Mean $\bar{Q}_1$, Q_2, Signed Q_2, and Q_3 Values and Kruskal-Wallis (K-W) Statistics for Data Sets for Grade 6

Data Set & Item Type	$\bar{Q}_1$	Q_2	Signed Q_2	Q_3
Data Set 9				
CTBS/U	19	17	7	.01
DMI	24	21	6	.00
Between	21	13	-6	-.03
K-W	1	29	215	334
Data Set 10				
Add	17	35	20	.02
Subtract	20	20	8	.01
Multiply	25	16	4	.01
Divide	23	18	10	.02
Between	21	14	-3	-.02
K-W	<1	14	117	173
Data Set 11				
Whole numbers	19	15	5	.01
Decimals	22	24	11	.01
Fractions	25	48	39	.05
Between	22	13	-6	-.02
K-W	4	28	301	297
Data Set 12				
Add whole	18	10	10	.04
Sub whole	21	28	27	.06
Mult whole	20	21	13	.03
Divide whole	20	19	16	.04
Add decimals	16	62	57	.08
Sub decimals	23	32	24	.04
Mult decimals	26	25	8	.01
Div decimals	35	37	29	.05
Add fractions	18	106	106	.16
Sub fractions	21	49	42	.06
Mult&Div frac	44	11	8	.02
Between	23	14	-3	-.02
K-W	5	47	256	305
Data Set 13				
CTBS/U	15	11	4	.01
DMI	20	15	4	.00
Between	18	11	-5	-.02
K-W	10	<1	382	557
Data Set 14				
Number sent	19	26	18	.03
Measurement	18	12	3	.01
Geometry	16	12	7	.02
Problem solv	20	11	3	.01
Numeration	15	19	8	.00
Number theory	21	87	77	.06
Sequences	22	134	134	.27
Ineq/Odd/Mult	17	9	9	.01
Between	18	11	-3	-.01
K-W	1	7	141	198
Data Set 15				
Greater than	10	89	89	.15
Fractions	25	87	83	.15
Estimation	20	591	591	.52
Seq/Ineq/Odd	21	60	58	.13
Conven Msrmt	11	86	86	.19
Place val DMI	10	784	784	.59
Expand notat	29	380	380	.49
Place val CTBS	11	64	64	.14
Other	18	11	0	-.00
Between	19	12	-4	-.02
K-W	20	21	180	290
Data Set 16				
CTBS/U	20	19	14	.03
DMI	18	15	7	.01
Between	19	13	-9	-.03
K-W	1	32	1176	1668
Data Set 17				
CTBS/U	16	13	8	.02
DMI	23	21	11	.01
Between	19	12	-9	-.04
K-W	<1	6	1110	1459

of local dependence that was evident for some item sets.

When a MC test and a MC&A test were analyzed together, there was evidence of substantial local dependence (as measured by signed Q_2 and Q_3), particularly for Grade 6 (see Data Sets 7, 8, 16, and 17). The Grade 6 mixed-content Kruskal-Wallis tests for signed Q_2 and Q_3 were the most significant tests of all those computed.

The Q_2 and Q_3 values were examined for items at the ends of tests. There was no evidence of local dependence among these items due to their position in the test.

Fit measures and item parameters. In order to determine whether local dependence appeared greater for items with similar item parameters, correlations were obtained between the absolute value of the difference between parameter values for each pair of items and signed Q_2 or Q_3. For signed Q_2 this correlation ranged from $-.05$ to $-.19$ for the $\hat{a}_i$

parameters, from .00 to −.18 for the $\hat{b}_i$ parameters, and from −.03 to −.28 for the $\hat{c}_i$ parameters. For Q_3 this correlation ranged from −.05 to −.20 for the $\hat{a}_i$ parameters, from .06 to −.11 for the $\hat{b}_i$ parameters, and from −.02 to −.40 for the $\hat{c}_i$ parameters.

Figure 1 shows the plot of the signed Q_2 values and the absolute differences between $\hat{b}_i$ parameters for Grade 6 CTBS/U-MC DMI-MC items. This plot (which represents a correlation of −.14) is typical of the plots for the other tests and parameters for signed Q_2. Figure 2 is the plot of the Q_3 values and the absolute differences between $\hat{b}_i$ parameters for the same items as those in Figure 1. As seen in a comparison of Figures 1 and 2, the signed Q_2 values tend to have a tighter plot than the Q_3 values; in both figures, however, high Q_2 or Q_3 values occurred only for items with similar parameters, but most item pairs that had similar parameters did not have high Q_2 or Q_3 values.

An analysis was done to see if the item sets that appeared to be locally dependent in Tables 6 and 7 had distinctive parameters. The items were sorted into sets as in Tables 6 and 7 and a Kruskal-Wallis test was used to test the hypothesis that the parameters for the items in the different sets all came from the same population. Tables 8 and 9 display the means of the parameters by sets and the values of the Kruskal-Wallis statistics. In many cases, but not all, item sets that showed substantial local dependence also showed $\hat{a}_i$ or $\hat{b}_i$ parameters that differed from items in other sets. For MC the most highly locally dependent item sets tended to be among the most difficult and most discriminating (see Data Sets 2, 3, 11, and 12). For MC&A the most highly locally dependent item sets were not necessarily extreme in terms of their difficulties or discriminations (see Data Sets 6, 14, and 15). The $\hat{c}_i$ parameters did not tend to show significant differences among item sets except for the Grade 6 tests that differed substantially in their numbers of answer choices (see Data Sets 9, 13, 16, and 17).

Trait estimates. As described in the Method section, the Grade 3 MC&A tests and the Grade 6 MC tests were divided into sets of items to obtain trait estimates and to examine how local dependence affects equatings. Table 10 displays the mean $\hat{a}_i$ and $\hat{b}_i$ values and the mean proportion correct for the items entering into each trait estimate. For the Grade 3 tests the CTBS/U items tended to be

Figure 1

Absolute Difference Between $\hat{b}_i$ Values for Each Item Pair Plotted Against Signed Q_2 for Grade 6 CTBS/U-MC DMI-MC Items

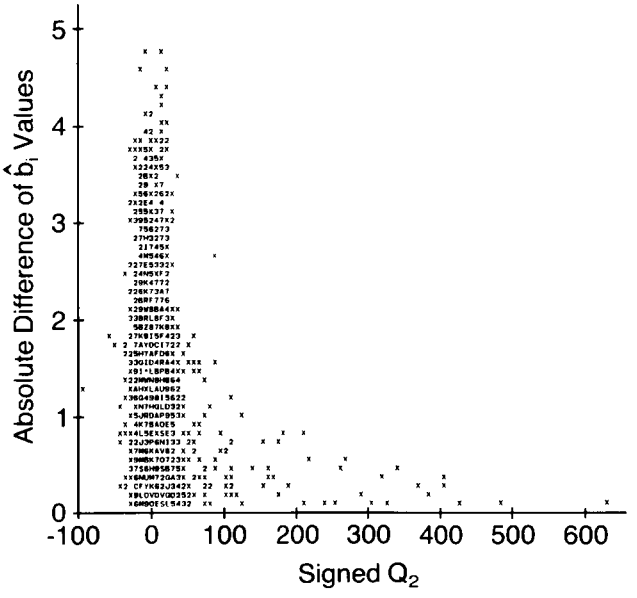
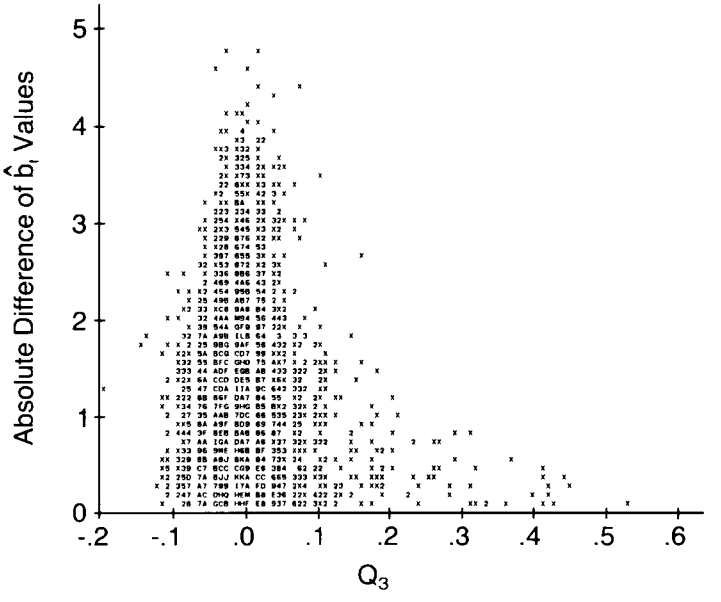


Figure 2
Absolute Difference Between $\hat{b}_i$ Values for Each Item Pair
Plotted Against Q_3 for Grade 6 CTBS/U-MC DMI-MC Items



more discriminating and difficult than the DMI items. The Locally Dependent items tended to be slightly more difficult than the Control items. For Grade 6 the CTBS/U and DMI items had similar difficulties and discriminations. Whole number computations tended to be easier than decimal computations, which tended to be easier than fractions computations. The difficulties of the Control items roughly followed the difficulties of the Whole Numbers, Decimals, and Fractions items. Although the Whole Number and Fractions sets (and the Control 1 and Control 3 sets) differed enough in difficulty to be considered to involve vertical equating, all the remaining equatings were horizontal equatings.

The results of the equatings are in Table 11. Compared to the equatings involving the Control tests at Grade 3, the CTBS/U and DMI equating showed a slightly higher standardized mean difference and a substantially lower correlation, higher SRMSD, and greater amount of local bias. The CTBS/U scores tended to be higher than the DMI scores among low scoring examinees and lower than the DMI scores among high scoring examinees. The equatings involving the Locally Dependent item set had the same standardized mean difference as the Control 1 to Control 2 equating but lower correlations and higher SRMSD values.

In Grade 6 the CTBS/U to DMI equating had about the same size standardized mean difference as the other equatings, except for the equatings involving Fractions. The correlation between CTBS/U scores and DMI scores was lower than those involving the Control items but higher than those involving the Fractions items. The Whole Numbers and Decimals equating had the same standardized mean difference but a lower correlation than the Control 1 to Control 2 equating. The Fractions equatings had higher standardized mean differences, lower correlations, and higher local bias than the Control 1 to Control 3 equating. An examination of plots of Fractions scores with other scores revealed that examinees who got very low scores on the Fractions items could get fairly high scores on the Whole Numbers and Decimals items.

Table 8

Mean $\hat{a}_i$, $\hat{b}_i$, and $\hat{c}_i$ Values and Kruskal-Wallis (K-W) Statistics for Data Sets for Grade 3

Data Set & Item Type	$\hat{a}_i$	$\hat{b}_i$	$\hat{c}_i$	Data Set & Item Type	$\hat{a}_i$	$\hat{b}_i$	$\hat{c}_i$
Data Set 1				Data Set 6			
CTBS/U	1.07	-.15	.13	Graph read	1.36	-1.40	.14
DMI	1.16	-.49	.11	Clock	.83	-.68	.14
K-W (1)	<1	<1	<1	Nearest ten	1.09	.33	.25
Data Set 2				Place value	1.02	.08	.17
Add	.86	-.94	.10	Number sent	.90	1.25	.23
Subtract	1.09	-.03	.11	Even numbers	1.22	.31	.22
Multiply	1.24	-.22	.17	Before	.60	-2.10	.14
Divide	1.42	.99	.11	Number line	.52	-.89	.14
K-W (3)	14[a]	11	12[a]	Multiply	.94	.85	.11
Data Set 3				Fractions	.89	1.47	.13
Add w/o rg	.68	-2.29	.10	Sequences	1.00	-.57	.17
Add with rg	.99	.11	.10	Shapes	.47	-2.30	.14
Sub w/o rg	.76	-1.06	.11	Other	.90	-1.00	.15
Sub with rg	1.52	1.31	.11	K-W (12)	19	33[a]	20
Mult by 5	1.47	-.35	.22	Data Set 7			
Other mult	1.13	-.16	.15	CTBS/U	.92	-.14	.11
Divide ÷	1.40	1.09	.09	DMI	.67	-1.22	.07
Divide $\sqrt{\ }$	1.45	.90	.14	K-W (1)	18[a]	8[a]	<1
K-W (7)	37[a]	42[a]	13	Data Set 8			
Data Set 4				CTBS/U	1.12	-.25	.19
CTBS/U	1.09	-.27	.17	DMI	1.08	-.48	.16
DMI	.71	-1.21	.14	K-W (1)	2	<1	2
K-W (1)	35[a]	4	2				
Data Set 5							
Count & match	.59	-4.06	.15				
Numeration	1.08	-.07	.20				
Number theory	.96	-.08	.17				
Measurement	.91	-.70	.14				
Geometry	.76	-1.03	.16				
Problem solv	1.12	-.91	.15				
Number sent	.80	-.23	.13				
Sequences	.84	-.46	.16				
K-W (7)	14	13	7				

Note. The degrees of freedom for the Kruskal-Wallis statistics appear in parentheses following "K-W".

[a] $p<.01$.

Discussion

Simulated Data

Properties of the fit measures. $\overline{Q}_1$ values had low correlations with Q_2 and Q_3 values, and the factors that cause misfit as measured by $\overline{Q}_1$ do not appear to include multidimensionality. Van den Wollenberg (1982) found a Rasch fit measure analogous to Q_1 not to be sensitive to multidimensionality, and Yen (1981) noted that Q_1 was not useful for determining when a two-parameter model was inappropriately applied to three-parameter data.

Table 9

Mean $\hat{a}_i$, $\hat{b}_i$, and $\hat{c}_i$ Values and Kruskal-Wallis (K-W) Statistics for Data Sets for Grade 6

Data Set & Item Type	$\hat{a}_i$	$\hat{b}_i$	$\hat{c}_i$	Data Set & Item Type	$\hat{a}_i$	$\hat{b}_i$	$\hat{c}_i$
Data Set 9				Data Set 14			
CTBS/U	.96	-.32	.16	Number sent	.95	-.09	.13
DMI	.87	-.31	.06	Measurement	.86	-.22	.15
K-W (1)	3	<1	48[a]	Geometry	.77	1.05	.13
Data Set 10				Problem solv	.95	-.06	.15
Add	1.00	-.33	.12	Numeration	.96	-.34	.17
Subtract	.99	-.55	.11	Number theory	1.09	.05	.13
Multiply	.68	-.33	.10	Sequences	.86	-1.17	.03
Divide	1.01	-.04	.10	Ineq/Odd/Mult	.77	-.28	.08
K-W (3)	13[a]	4	<1	K-W (7)	8	18	8
Data Set 11				Data Set 15			
Whole num	.90	-.75	.11	Greater than	1.27	.73	.22
Decimals	.84	-.26	.10	Fractions	.86	-.04	.07
Fractions	1.13	.55	.13	Estimation	1.42	.45	.10
K-W (2)	2	15[a]	3	Seq/Ineq/Odd	.84	-.81	.05
Data Set 12				Conven Msrmnt	.95	1.01	.09
Add whole	.47	-1.44	.11	Place val DMI	.59	.54	.08
Sub whole	.76	-1.20	.09	Expand notat	1.09	-.81	.07
Mult whole	.76	-1.01	.10	Place vl CTBS	.88	-.69	.13
Divide whole	1.12	-.27	.11	Other	.89	.04	.16
Add decimal	.96	-.48	.12	K-W (8)	16	12	18
Sub decimal	.87	-.78	.10	Data Set 16			
Mult decimal	.68	.20	.10	CTBS/U	.85	-.37	.14
Div decimal	.75	.50	.08	DMI	.90	.30	.10
Add fraction	1.34	.54	.13	K-W (1)	<1	9[a]	10[a]
Sub fraction	1.41	.49	.14	Data Set 17			
Mult&Div fra	.39	.65	.10	CTBS/U	.93	-.21	.25
K-W (10)	39[a]	30[a]	6	DMI	.83	-.33	.05
Data Set 13				K-W (1)	5	<1	69[a]
CTBS/U	.94	-.28	.21				
DMI	.89	.26	.08				
K-W (1)	<1	7[a]	58[a]				

Note. The degrees of freedom for the Kruskal-Wallis statistics appear in parentheses following "K-W".

[a] $p<.01$.

Thus, although Q_1 can be useful in identifying items that have unexpected item characteristic curves, Q_1 cannot be relied upon as a complete fit measure.

Signed Q_2 and Q_3 produced essentially the same results. It is advantageous to have both signed and unsigned statistics that can be compared, as with signed Q_2 and Q_2; these statistics also have magnitudes that are easy to analyze. The small values of the Q_3 statistics make them more difficult to evaluate subjectively, but the Q_3 values are readily interpreted as correlations. The Q_2 values will increase as a function of sample size, while the Q_3 values will not.

In order to understand the differences in the results for the three two-dimensional simulations, it is necessary to consider the signed Q_2 or Q_3 values

Table 10

Mean Item Difficulties (p_i), $\hat{a}_i$, and $\hat{b}_i$ Values for Items Entering Into Trait Estimates

Grade	Content	Trait	$\hat{a}_i$	$\hat{b}_i$	p_i	No. of Items
3	MC&A					
		CTBS/U	1.09	-.27	.63	40
		DMI[a]	.71	-.76	.68	40
		Loc.Dep.	.91	-.23	.59	27
		Control 1	.91	-.65	.66	26
		Control 2	.94	-.66	.66	25
6	MC					
		CTBS/U	.96	-.32	.62	40
		DMI	.87	-.31	.58	48
		Whole num.	.90	-.75	.65	36
		Decimals	.84	-.26	.57	36
		Fractions	1.13	.55	.41	16
		Control 1	.82	-.57	.62	36
		Control 2	1.04	.11	.51	36
		Control 3	1.03	.43	.47	16

[a]Two items with $\hat{b}_i$<-5 were not included in the mean $\hat{b}_i$.

Table 11

Comparisons of Trait Estimates

$\hat{\theta}_1$	$\hat{\theta}_2$	r	$\frac{(\bar{\hat{\theta}}_1 - \bar{\hat{\theta}}_2)}{\bar{S}}$	$\frac{S_1}{S_2}$	SRMSD	Local Bias
Grade 3 MC&A						
CTBS/U	DMI	.74	-.05	.92	.73	.07
Loc.Dep.	Control 1	.81	.00	1.05	.62	.03
Loc.Dep.	Control 2	.80	.00	1.01	.63	.01
Control 1	Control 2	.85	.00	.96	.56	.03
Grade 6 MC						
CTBS/U	DMI	.76	-.05	1.07	.70	.06
Whole num.	Decimals	.77	.02	.99	.68	.03
Control 1	Control 2	.89	.02	.96	.48	.04
Whole num.	Fractions	.65	.09	.86	.85	.17
Decimals	Fractions	.64	.08	.87	.86	.16
Control 1	Control 3	.83	.04	.89	.60	.09

and to reexamine the simulation conditions in Table 2. In Configuration 1, both underlying traits influenced every item. The correlation between $\hat{a}_i$ and the sum of the a_{is} values used in generating the data was .70. It appears that the unidimensional model used a combination of the two underlying

traits as the unidimensional trait. When a combination of traits is used to sort examinees into cells, local negative dependence appears. (This effect is described in the earlier section that gives the motivation for the development of signed Q_2.) Thus, for Configuration 1, the vast majority of item pairs had negative signed Q_2 and Q_3 values. The items did not differ a great deal in terms of their relative a_{is} values for the two underlying traits, and the within-set and between-set signed Q_2 and Q_3 values did not appear significantly different.

For Configuration 2, the correlation between the $\hat{a}_i$ values and the $a_{i1} + a_{i2}$ values was .89. Again, the unidimensional model used a combination of the underlying traits to define the unidimensional trait. However, in this configuration only the items in Set 1 were influenced by both underlying traits, and these Set 1 items all showed negative local dependence. The items in Sets 2 and 3 were influenced by only one of the underlying traits, and they showed positive local dependence; the between-set local dependence was largely negative.

For Configuration 3, the second trait was a very weak trait compared to the first trait, and the unidimensional trait used to sort examinees into cells was largely Trait 1. The items in Set 2, which were influenced only by the first underlying trait, showed a low amount of negative local dependence, as would be consistent with part-whole contamination discussed earlier. The items in Set 1, which were influenced by both underlying traits, showed positive local dependence.

These results for the two-dimensional simulations can be summarized as follows. If a combination of two underlying traits is used as the unidimensional trait, then items that are influenced by both underlying traits will show negative local dependence and items that are influenced by only one underlying trait will show positive local dependence. If only one of the underlying traits is used as the unidimensional trait, then items that are influenced only by that underlying trait will show slight negative local dependence due to part-whole contamination and items that are influenced by both underlying traits will show positive local dependence.

Real Data

Properties of the fit measures. The real data sets tended to have positive local dependence within sets and negative local dependence between sets. These results appear most consistent with a multidimensional situation in which there are important, different underlying traits influencing the items within the different item sets and the unidimensional estimated trait is a combination of these underlying traits.

The overall significance tests involving Q_2 or Q_3 were not particularly useful by themselves. It was important to examine the direction of the local dependence both within and between sets of items that were suspected of being multidimensional. It was also important to keep in mind that signed Q_2 and Q_3 will have slightly negative values for unidimensional data, and that any positive correlations that are found may be underestimates of the strength of the positive relationships between the items.

For the real data, Q_2 and Q_3 identified item sets as being locally dependent that a priori would appear to be so. For the MC tests the locally dependent item sets corresponded to category objectives that were created through a logical analysis of the item content. For the MC&A tests the category objectives did appear to be locally dependent, but item sets that were much more locally dependent than the category objectives were identified through an examination of the Q_2 values. (These same locally dependent item sets would be identified through an examination of Q_3 values.) Thus, the construct validity of Q_2 and Q_3 was enhanced by the fact that every item set that was identified as having high Q_2 or Q_3 values had a logical association of content among the items in the set. Q_2 and Q_3, therefore, appeared useful for sorting items into sets measuring different dimensions. This information can be valuable in both test construction and in test scoring.

Q_2 and Q_3 also distinguished between items that might be expected to be locally dependent. For example, for Data Set 15 CTBS/U place value items and DMI place value items were locally dependent within each set but not across sets. A closer ex-

amination of the items revealed that while the item stems had the same format, the types of answer choices were quite different for the two tests; this difference in the answer choices also affected the difficulty and discrimination of the items. Thus, Q_2 or Q_3 could be useful in item tryouts to distinguish among item formats and to help identify the most useful ones.

Properties of locally dependent items. The items with high Q_2 or Q_3 values tended to have similar item parameters, but the items with similar parameters did not necessarily have high Q_2 or Q_3 values. High values of Q_2 or Q_3 were not artifacts of similarity of item parameters. Similarity of item parameters is a necessary, but not a sufficient condition for high Q_2 or Q_3 values. Items need to share one or more unique traits or dimensions in addition to having similar item parameters in order to have high Q_2 or Q_3 values. The Q_2 values are a function of the within-cell ϕ_{ijr} values, and the ϕ_{ijr} values are restricted in magnitude unless items have similar within-cell p values; an analogous effect occurs with Q_3. Thus, locally dependent items that do not have similar item parameters are unlikely to have high Q_2 or Q_3 values; it is not known how often such items occur.

The MC items that were highly locally dependent tended to have high discriminations and difficulties, but the MC&A items did not. The MC items tended to be naturally hierarchical in nature. For example, students learn addition without regrouping before they learn addition with regrouping, and they tend to learn some addition before they learn subtraction or multiplication. The MC item sets tend to be correlated and cumulative. For example, to do some division items, examinees need to be able to add, subtract, and multiply. Given this structure of the MC items, it is not surprising that the more complex items appeared to be locally dependent and have high discriminations and difficulties.

The highly locally dependent MC&A items tended to be "pockets" of separate, fairly independent concepts. Although there would tend to be some correlation among these concepts due to individual differences in speed of progress through the curriculum, many of the MC&A concepts could be learned fairly independently of the others, and these concepts were not strongly cumulative in nature. Unlike the MC items, there is no reason to expect the MC&A locally dependent items to be highly discriminating or difficult.

Unidimensional trait estimates. It can be hypothesized that when the unidimensional three-parameter model analyzes test data generated from several correlated underlying traits, the unidimensional model uses a combination of the underlying traits as its unidimensional trait. Yen (1984) presented derivations and simulated data that support this hypothesis. McKinley (1983) reported supporting simulation data for the two-parameter model. The fact that local dependence between sets was negative also supports the hypothesis that a combination of traits was used to create the unidimensional trait estimate.

The hypothesis that the unidimensional trait is a combination of correlated underlying traits is consistent with the fact that with the MC data the more complex items tended to have higher discriminations; it is also consistent with the fact that for Grade 3 MC&A, the CTBS/U items had higher average discriminations than the DMI items. For the Grade 3 CTBS/U MC&A test, examinees must read text associated with each item. For the Grade 3 DMI MC&A items, any associated text is read to the examinees. Thus, it can be hypothesized that the unidimensional model chooses a combination of mathematics and reading skills to define its trait; because the CTBS/U items involve both traits, they tend to have higher discriminations.

It can also be hypothesized that when a test involves independent traits that influence only a few items, such traits are ignored in the definition of the unidimensional three-parameter trait. Reckase (1979) reported results that support this hypothesis. The present study did not include simulations with independent traits, so it does not assist in the evaluation of this hypothesis; but the hypothesis is consistent with the results for the MC&A tests where the pockets of locally dependent items did not have particularly high discriminations.

Equating tests in the presence of local depen-

dence. Trait estimates based on items measuring different dimensions had lower correlations than trait estimates that had items that shared dimensions. However, except for the Grade 6 equatings involving the Fractions items, there were not large systematic errors of equating due to the attempted equating of items measuring different dimensions. As might be expected if fractions are frequently taught after whole numbers and decimals, examinees could get moderately high Whole Numbers or Decimals scores while having very low Fractions scores. Thus, it appears that while substantial unsystematic errors of equating are to be expected from the equating of tests involving collections of different dimensions, substantial systematic errors of equating might not be expected unless the two tests measure quite different dimensions that are taught sequentially and that differ in difficulty. Only a handful of trait estimates have been compared here, and a wide variety of conditions can occur with multidimensional tests. It would be prudent to cross-validate any equatings that appear to involve multidimensional tests.

When maximum likelihood scoring of item responses is used with the three-parameter model, item discriminations are used in the item weights. Thus, when a test measures several correlated dimensions, items that measure more of these dimensions will tend, all else being equal, to have higher discriminations and to be given more weight in the test scoring. An alternative scoring procedure is to use unit item weights, so that a number-correct score is converted to a trait estimate (Yen, in press). If the test user wants to measure several dimensions with one score, IRT scaling combined with differential item weighting is a potentially efficient method of test scoring. The desirability of this procedure is a function of whether the additional dimensions are legitimate aspects of the behavior that the test user wants to measure. In any multidimensional testing situation it would be important to review the item discriminations and fit statistics such as Q_2 or Q_3 to help determine what dimensions are influencing test performance. Whether or not items are differentially weighted and whether or not IRT is used, when a test measures several dimensions, examinees' scores will be influenced by all these dimensions. Thus, unsystematic and systematic errors of equating might be expected from any scaling and equating procedures that are applied to some multidimensional tests.

References

CTB/McGraw-Hill. (1981). *Comprehensive Tests of Basic Skills, Form U*. Monterey CA: Author.

CTB/McGraw-Hill. (1982a). *Comprehensive Tests of Basic Skills, Forms U & V, Preliminary Technical Report*. Monterey CA: Author.

CTB/McGraw-Hill. (1982b). *Comprehensive Tests of Basic Skills, Forms U & V, Test Coordinator's Handbook*. Monterey CA: Author.

Doody-Bogan, E., & Yen, W. M. (1983, April). *Detecting multidimensionality and examining its effects on vertical equating with the three-parameter logistic model*. Paper presented at the meeting of the American Educational Research Association, Montreal.

Gessel, J. (1975a). *Diagnostic Mathematics Inventory*. Monterey CA: CTB/McGraw-Hill.

Gessel, J. (1975b). *Diagnostic Mathematics Inventory Teacher's Guide*. Monterey CA: CTB/McGraw-Hill.

Hays, W. L. (1973). *Statistics for the social sciences*. San Francisco: Holt, Rinehart, & Winston.

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

Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13*, 517–549.

McKinley, R. L. (1983, April). *A multidimensional extension of the two-parameter logistic latent trait model*. Paper presented at the meeting of the National Council on Measurement in Education, Montreal.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*, 207–230.

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

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

Traub, R. E. (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.),

Applications of Item Response Theory (Monograph). British Columbia: Educational Research Institute of British Columbia.

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

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST User's Guide*. Princeton NJ: Educational Testing Service.

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

Yen, W. M. (1983a). Tau equivalence and equipercentile equating. *Psychometrika, 48,* 353–370.

Yen, W. M. (1983b). Use of the three-parameter logistic model in the development of a standardized achievement test. In R. K. Hambleton (Ed.), *Applications of Item Response Theory* (Monograph). British Columbia: Educational Research Institute of British Columbia.

Yen, W. M. (1984, June). *Increasing item complexity: A possible cause of scale shrinkage for unidimensional item response theory*. Paper presented at the meeting of the Psychometric Society, Santa Barbara CA.

Yen, W. M. (in press). Obtaining maximum likelihood trait estimates from number-correct scores for the three-parameter logistic model. *Journal of Educational Measurement*.

Author's Address

Send requests for reprints or further information to Wendy M. Yen, CTB/McGraw-Hill, 2500 Garden Road, Monterey CA 93940, U.S.A.