Scaling Performance Assessments: Strategies for Managing Local Item Dependence

Wendy M. Yen
CTB Macmillan/McGraw-Hill

Performance assessments appear on a priori grounds to be likely to produce far more local item dependence (LID) than that produced in the use of traditional multiple-choice tests. This article (a) defines local item independence, (b) presents a compendium of causes of LID, (c) discusses some of LID's practical measurement implications, (d) details some empirical results for both performance assessments and multiple-choice tests, and (e) suggests some strategies for managing LID in order to avoid negative measurement consequences.

Performance assessments require qualitatively different performance of students than do multiple-choice tests. One of the ways in which performance assessments differ from traditional multiple-choice tests is the expected level of local item dependence (LID). The items in traditional multiple-choice tests are usually carefully designed to be independent of one another. That is, the success on one item is not influenced by success on another. Multiple-choice items are not chained, and theoretically they could be presented to the student in any order without affecting the item difficulty.

In contrast, for performance assessments a setting is established, and students can be asked to make multiple responses to directions or questions related to that setting. For example, consider a grade 5 test in language arts. The setting might be established with a short story. The student is asked to contrast two characters in the story, provide and defend an alternative story ending, and relate events in the story to a personal experience. In mathematics at grade 8, the setting could be a school holding a bake sale to raise money. Given appropriate information, the student is asked to figure the costs for making a given amount of baked items, calculate the profit margin per item, decide whether a particular change in the proportion of cookies and cakes would be likely to increase profits, and explain the reasoning underlying this decision.

Such performance assessments appear likely to produce far more LID than that produced in the use of traditional multiple-choice tests. The following sections define local item independence and describe possible causes of violations of independence. Some of LID's practical measurement implications are then discussed. Empirical results related to LID for both performance assessments and multiple-choice tests are presented, and finally some strate-

The most interesting data in this article were obtained from the Maryland School Performance Assessment Program, and the author is grateful to the Maryland State Department of Education for its contribution of that data to this study.
gies are suggested for managing LID in order to avoid negative measurement consequences.

A note on terminology is appropriate here. The questions or directions given to students in a performance assessment are very different from those in multiple-choice tests. Performance assessment scores are also dependent on the rule or rubric used to grade each response. For the sake of simplicity of terminology, each prompt, question, or set of directions to which a student responds, along with its associated scoring rule, will be called collectively an item. The student product will be called the response. In cases where confusion will not arise, the score assigned by a rater to a student product will also be called the response.

**Definition of Local Item Independence**

The concept of conditional independence of item scores is used, in various forms, in classical true score theory, latent class analysis, factor analysis, and item response theory (Lord & Novick, 1968). In classical true score theory, the assumption is made that errors of measurement are uncorrelated given the examinee's true score. In item response theory, when a pair of items is locally independent, the conditional probability, given the student's ability level, θ, of obtaining any pair of scores on these items is the product of the probabilities for the separate items:

\[ P(X_1 = x_1 \text{ and } X_2 = x_2 | θ) = P(X_1 = x_1 | θ) \cdot P(X_2 = x_2 | θ). \] (1)

The assumption being made is that the true score or trait value is providing all the relevant information about the student's performance and that the contribution of each item to the test can be evaluated independently of all other items.

The direction, positive or negative, of LID refers to the conditional correlation of item scores. Positive LID means that, if a student performs higher (or lower) than expectation on one item, he or she probably will perform higher (or lower) than expectation on the other. The expectation is based on overall test performance. Negative LID means that, if a student performs unusually well on one item, he or she probably will perform unusually poorly on the other.

If items scores are locally dependent, then, in factor analysis terminology, they will have nonzero residual correlations after removal of the first factor. That is, imagine that a factor analysis is broken down by steps. The first factor is defined and removed from the data; that step is roughly equivalent to conditioning on θ in Equation 1. If the residual item correlations are all zero, then there is no local item dependence, and no more than one factor will be defined. However, if there are sets of items that have unusually large residual correlations, then those items will define a second factor.

**Causes of Local Item Dependence**

This section outlines some of the possible causes of LID. The basic principle involved in producing LID is that there is an additional factor that consistently affects the performance of some students on some items to a greater extent than others. In a factor analysis, such items would be likely to load on a second (or higher) factor. These additional factors may be desirable because they involve important dimensions of the behavior being measured, or they may be undesirable nuisance factors. Constant effects, even if they are unintended, do not produce LID if they apply equally to all students and/or to all items.

**External assistance or interference.** If, for some items, some students receive effective assistance from a teacher or fellow student, they will do unusually well on the facilitated items, producing LID. In contrast, an effect that lowers scores and produces LID is external interference, such as classroom disruption, faulty materials or administration procedures, or inaccurate information from a teacher or peer that affects some students and/or some items.

**Speededness.** If there is insufficient time for a substantial proportion of students to finish a test, item responses near the end of the test will be positively, locally dependent. If time management is an important factor in test taking, then substantial negative LID can also occur; if the student chooses to spend time on one section of the test rather than another, the student will tend to have unusually high scores on the section on which time was devoted and unusually low scores on the section that received less attention.

**Fatigue.** Passages or tasks tend to be more difficult when they appear at the end of a demanding test than if they appear at the beginning. Shared fatigue or lowered motivation among items in such passages can produce positive LID.

**Practice.** If item performance tends to improve with practice or exposure to items, then LID can occur. If the items are relatively homogeneous in content, local dependence will be related solely to the relative position of the items in the test. An unusual format will, for some students, "take some getting used to," producing LID.

**Item or response format.** Items can be designed to measure the same content while using different formats. Responses to items can vary, involving either selected or constructed responses. Constructed responses can vary in terms of their length or type, as when a student can respond by writing a story, drawing a picture, or building a model. These variations can all produce LID.

**Passage dependence.** If several items are attached to the same passage or setting, then LID can occur. This LID can be produced by a student's unusual level of interest or background knowledge about the passage or by the fact that information used to answer different items is interrelated in the passage.

**Item chaining.** If items are organized in steps, then knowing the answer to one item increases the chances of a student's knowing the answer to the next one. While item chaining has long been an anathema to multiple-choice tests, it is often seen as desirable in performance assessments because it models real life situations.

**Explanation of previous answer.** A special type of item chaining can be found in mathematics performance assessments. In one item, a student is asked to calculate an answer, draw a conclusion, or make a decision. In a subsequent item, the student is asked to explain the reasoning or process used to produce the previous answer. The explanation, while providing additional information...
about student performance, would be expected to be highly locally dependent on the preceding item.

**Scoring rubrics or raters.** The items in a performance assessment can be scored by a variety of rubrics or rules. Items that are scored with the same rubric may display LID, either because they measure common skills in the students or because they place special demands on the raters. LID can also be produced if a test is divided into sections for scoring purposes, if a variety of raters provide ratings, and if there are uncontrolled rater effects, such as halo or leniency effects.

**Content, knowledge, and abilities.** Achievement tests typically involve items that measure a range of content. Items can display LID if they measure unique content. For example, in a grade 3 mathematics concepts test, items measuring clock reading can tend to be locally dependent, as can items measuring fractions. When performance is differentially affected by exposure in the curriculum or in everyday life (i.e., opportunity to learn), items with the same exposure can exhibit LID (Yen, 1984a). Items that display differential item functioning frequently would qualify under this category and can exhibit LID.

### Measurement Implications of Local Item Dependence

One of the biggest differences between multiple-choice tests and performance assessments is the degree of focus of each item. With multiple-choice educational achievement tests, particularly those developed in the 1970s and 1980s, great attention was paid to developing discrete items closely tied to objective structures and separating performance into clearly defined pieces. The philosophy behind performance assessment is the diametric opposite, rejecting artificial dissection of performance and embracing the measurement of behavior as an intact whole.

No test can perfectly satisfy every goal for its construction; trade-offs must be made. For example, the goals of reliability and broad content coverage must be balanced with the goal of meeting limits in testing time. To meet the goal of authenticity, it may be seen as necessary to have “large” (i.e., time-consuming), “rich” items that are dependent. In order to evaluate whether it is worthwhile to use independent items, the goals of the testing program and the practical implications of LID on these goals must be specified. Only then can it be determined whether achieving local item independence is an unimportant psychometric nicety that can be ignored or whether it has important practical effects that merit attention.

Some typical goals of educational achievement tests will be briefly described, and the effects on LID on meeting those goals will be discussed.

If the only goal of a performance assessment is the one-time measurement of behavior as a whole, then it is possible for one item to be sufficient to meet that goal. However, if the following additional goal is part of the assessment, additional, independent items are necessary:

1. scores that differentiate a student’s relative performance on different outcomes.

If the following goals exist, then it is likely that the more independent items that are used within a test form, the more closely these goals will be met:

2. scores that can be validly generalized beyond the specific context of the test form,

3. scores that are sufficiently reliable to make decisions about individual students with some acceptable amount of error in these decisions,

4. calibration of items for an item bank or task bank from which items with known measurement characteristics can be drawn, and

5. scores that can be equated or compared over forms.

The specific effects of LID on these goals can be described.

**Outcome or objective scores.** It is common in educational achievement tests for diagnostic scores to be reported based on subsets of items that measure different outcomes. If items are locally dependent across those categories, it can be difficult for performance on such outcomes to be evaluated separately. Validity. An essential aspect of validity is that a score be the foundation of appropriate decisions. These decisions may be about individual students or about groups, such as schools or states. Typically, these decisions are broad based, and the desire is to draw appropriate conclusions that cover a variety of situations.

Two factors that affect the usefulness of test scores for making decisions are the types of behaviors tested and the breadth of situations in which those behaviors are sampled. Greater authenticity is possible with performance assessments, but, in order for their scores to generalize to a variety of real life situations, they must be based on multiple samples of behavior. The greater the independence of the samples of behavior, and the greater the variety of contexts in which they are obtained, the greater the opportunity to produce a score that will generalize over the variety of real life behaviors that are typically of interest. When LID occurs, it can mean that the multiple observations of behavior are not covering as wide a range of behavior as intended.

**Test information and standard errors of measurement.** If a total score is based on multiple observations, the greater the number of observations, the greater the accuracy of the total score. However, multiple dependent observations will not produce as much increase in accuracy as multiple independent observations.

One way to look at the effect of LID on measurement error is to consider a simple classical true score theory model, where the score on the $j$th item is $X_j = T + E_j$ and the total test score based on $n$ items is

$$Y = \sum_{j=1}^{n} X_j$$

(2)

The measurement error in $Y$ is

$$SEM^2 (Y) = \sum_{j=1}^{n} \sigma^2 (E_j) + \sum_{j \neq k} \text{Cov} (E_j, E_k).$$

(3)
It is typically assumed that $\text{Cov}(E_j, E_j') = 0$, so that

$$SEM^2(Y) = \sum_{j=1}^{n} \sigma^2(E_j). \tag{4}$$

However, if in fact measurement error is positively correlated over items, the actual measurement error in the total test score will be greater than that described in Equation 4.

Within an item response theory (IRT) framework, the standard error of measurement ($SEM$) of the trait estimate is the reciprocal of the square root of the test information. Test information for optimal trait estimates is obtained by summing the information of the items contributing to the test score (Lord, 1980, chap. 5):

$$I(\hat{\theta}) = \sum_{j=1}^{n} I_j(\hat{\theta}) \tag{5}$$

The summation of the item information functions is justified by the assumption of local independence of the items. If in fact items are locally dependent, then Equation 5 will overstate the amount of independent information added by each item.

The $SEM$ is frequently used to construct confidence intervals for scores, so that users will be aware of the magnitude of variation of scores due to measurement error. Test information can also be used to estimate the number of items needed to reach a particular level of score accuracy. In computerized adaptive testing, it is common for items to be administered until a preset level of score accuracy has been reached. Misspecification of test information can affect all these applications.

**Item trace lines.** On each item, the IRT models relate the probability of each possible item score to the students’ ability levels or trait values ($\theta$). These descriptions are called trace lines. If an appropriate model is used, it typically can accurately describe the observed trace lines, regardless of whether or not LID occurs. In other words, fit measures that examine the match of predictions to observed trace lines are not affected by LID (Yen, 1984a). What LID affects is (a) the behavior of the observed trace lines themselves for individual items and (b) the probability of combinations of scores on items.

When positive LID occurs, it increases the strength of the relationship between some items and thereby can increase the strength of the relationship between an item and the total test score. This effect can produce higher item discriminations for LID items (Masters, 1988). This increase occurs in the observed trace lines and is reflected in the model predictions and parameters. If item banking is being used, and LID items are separated in subsequent test forms, then the predicted discrimination for such items can be inaccurately high. If items with the highest discriminations are chosen for use, either in a new intact test form or in computer adaptive testing, LID can produce inappropriate choices. It should be noted that the same effect will be produced by LID within a classical model in which items with the highest point biserials are chosen.

When a pair of items is locally independent, the probability of a student’s obtaining any pair of scores on these items is the product of the probabilities for the separate items as described in Equation 1. If in fact the items are not independent, Equation 1 will not be accurate. An application that would make use of these joint probabilities is predictions of distributions of raw scores for populations of students that have not actually taken a particular set of items. Such an application occurs but is not commonplace. A frequently used, essential application of trace lines is the development of scoring and equating tables.

**Trait estimates and equating.** In IRT, the item parameters relate scores on every item to the trait value. This relationship is used for converting a student’s item scores to a trait estimate. In optimal scoring, which produces a trait estimate with the lowest possible $SEM$, the student’s pattern of item responses is considered; in the optimal scoring, each item score is multiplied by a weight, which is a function of the item parameters. When it is unlikely that a student can get an item correct through guessing, as it is for constructed response tests, the optimal weights are proportional to the item discriminations (Lord, 1980). It is also possible to base trait estimates on unweighted raw scores; for multiple-choice tests, the unweighted raw score is the number-correct score while, for performance assessments, the unweighted raw score is the sum of the item scores.

Optimal scoring weights can be affected by LID. If LID increases an item’s discrimination and its optimal scoring weight, and the LID is due to a nuisance factor, such as interference, then the optimal weight is not desirable. Also, if the greater weight is based on LID produced by a pair of items being administered in the same form, the weight may not be appropriate if the items are separately administered. On the other hand, if the LID is due to better item quality or due to the items’ measuring additional important abilities or achievements, then the greater weight is desirable.

Every trait value can be related to an expected raw score, be it weighted or unweighted. For example, let the total unweighted raw score on a test be defined by Equation 2. It can be assumed that the $j$th item has $m_j$ possible scores. The raw score expected for each trait value, also known as the characteristic function, is the following:

$$E(Y|\theta) = \sum_{j=1}^{n} \sum_{k=1}^{m_j} x_{jk} P(X_j = x_{jk}|\theta). \tag{6}$$

Equation 6, or an analogous one based on weighted raw scores, can be used to produce a scoring table for transforming a raw score to a trait estimate. Thus, the accuracy of such equations is of great importance in the equating of test forms. It is important to note that item independence is not assumed in Equation 6. What is needed for accuracy of the equating is that the observed item trace lines are accurately described by the model. Thus, LID is not
expected to affect the accuracy of such tables unless item parameters are estimated for items in one test configuration and used for those items in a different configuration that has significantly different LID.

Empirical Results for Local Item Dependence

This section presents results related to a subset of the issues related to LID:
1. the extent of within-passage LID for performance assessments versus multiple-choice tests;
2. the extent of LID displayed by follow-up mathematics items that ask the student to explain how or why a previous answer was given;
3. the effect of an inappropriate assumption of item independence on:
   (a) model descriptions of standard errors of measurement,
   (b) item discriminations, and
   (c) item trace lines and characteristic functions.

These analyses were based on the following data sets and analysis procedures.

Data

Multiple-Choice Tests

Data from the Comprehensive Tests of Basic Skills, Fourth Edition (CTBS/4; CTB Macmillan/McGraw-Hill, 1989) were used. Randomly selected student response vectors were taken from a spring national standardization at grades 3, 5, and 8. Reading Comprehension (RC) and Mathematics Concepts and Applications (MC&A) tests were analyzed. All items had four answer choices. Omitted responses were given a score of 0. Table 1 presents the sample sizes and numbers of items in these analyses.

Performance Assessments

Data were taken from the Maryland School Performance Assessment Program (MSPAP), which involved the census testing of Maryland students in grades 3, 5, and 8 in May 1991. The data analyzed here were based on randomly selected samples of the students administered two of the (five or six) Language Arts forms and two of the (five) Mathematics forms used in each grade.

| TABLE 1 |
| Sample Sizes (N) and Numbers of Items (n): CTBS/4 |

<table>
<thead>
<tr>
<th>Grade</th>
<th>Reading Comp.</th>
<th>Math C &amp; A</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>n</td>
</tr>
<tr>
<td>3</td>
<td>2163</td>
<td>36</td>
</tr>
<tr>
<td>5</td>
<td>2165</td>
<td>40</td>
</tr>
<tr>
<td>8</td>
<td>1064</td>
<td>40</td>
</tr>
</tbody>
</table>

Student response books were graded by Maryland teachers according to detailed scoring rules and rubrics under the supervision of personnel from the Maryland State Department of Education and CTB Macmillan/McGraw-Hill. The lowest score for each item was 0, and the maximum possible score (MPS) varied from 1 to 3. Omitted item responses for sessions where a student was present were given a score of 0. For Reading, approximately 20% of the items had an MPS of 1, 60% had an MPS of 2, and 20% had an MPS of 3. For Mathematics, about 61% of the items had an MPS of 1, 33% had an MPS of 2, and 6% had an MPS of 3. More detailed information about the scoring rules is available from Fitzpatrick, Ercikan, and Ferrara (1992).

The Language Arts forms involved items measuring Reading, Writing, and Language Usage; only the Reading (RD) scale is examined here. All the Mathematics items in each form were placed on a Math Total (MT) scale; in addition, two scales involving overlapping sets of items were examined: Math Content (MC) and Math Process (MP). A student did not receive a scale score if he or she was absent for any of the testing sessions contributing to that scale.

For the Language Arts performance assessments, testing occurred in five sessions, spread out over 5 days, with 1.0 or 1.5 hours of testing per session. In each session, students read one or more short stories or articles and wrote several short answers or one extended response, depending on the session. Reading responses were obtained in four of the five sessions.

For the sake of simplicity of terminology, the short stories or articles that the students read are called passages. These passages measured the outcomes Reading for Information, Reading for Literary Experience, or Reading to Perform a Task. (Reading to Perform a Task did not involve actually performing the task.)

The Mathematics assessment was administered in three 1-hour sessions on 3 school days preceding (Grade 8) or following (Grades 3 and 5) the Language Arts assessment. For the forms examined in this study, in each session students provided responses to three tasks related to different themes. Within a session, the tasks were not separately timed, and students managed the time on their own.

The description of the scenario and data for each mathematics task typically took a half page. The students responded to several items related to that scenario. Additional information was often provided for later items. There were 3 to 11 items per task, with the typical number of items per task being 5 or 6. For the sake of efficiency of terminology, math items related to the same task are described as being related to the same passage.

The Math tests measured 13 outcomes, 9 of which were classified as Content outcomes and four as Process outcomes. An average of 68% of the items measured only Content outcomes; 36% measured only Process outcomes, and 20% measured both. It was common for a Process item to follow a Content item and ask the student to explain the reasoning or process behind the answer to that preceding Content item.

Table 2 displays the numbers of students and items involved in the performance assessment analyses.
Yen Scaling Performance Assessments

Sample Sizes

<table>
<thead>
<tr>
<th>TABLE 2</th>
<th>Sample Sizes and Numbers of Items: Performance Assessments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reading</strong></td>
<td><strong>Mathematics</strong></td>
</tr>
<tr>
<td><strong>Form</strong></td>
<td><strong>Number of Items</strong></td>
</tr>
<tr>
<td><strong>Grade 3</strong></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>5514</td>
</tr>
<tr>
<td>36</td>
<td>5889</td>
</tr>
<tr>
<td><strong>Grade 5</strong></td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>6728</td>
</tr>
<tr>
<td>52</td>
<td>6098</td>
</tr>
<tr>
<td><strong>Grade 8</strong></td>
<td></td>
</tr>
<tr>
<td>81</td>
<td>4226</td>
</tr>
<tr>
<td>82</td>
<td>4417</td>
</tr>
</tbody>
</table>

**Calibration Models**

**Multiple-Choice Tests**

The three-parameter logistic (3PL) model (Lord, 1980) was used for multiple-choice items. Maximum likelihood parameter estimates were obtained using the program PARMATE (Burket, 1990).

**Performance Assessments**

For an item with \( m_j \) levels, the item responses were integers ranging from 0 to \( m_j - 1 \). A two-parameter partial credit (2PPC) model was used to scale the item responses to performance assessments. The probability (trace line) of a student with ability \( \theta \) having a score at the \( k \)th level of the \( j \)th item is

\[
P_{jk}(\theta) = P(X_j = k - 1 | \theta) = \frac{\exp(z_{jk})}{\sum_{k=1}^{m_j} \exp(z_{jk})}
\]

where

\[
z_{jk} = a_{jk} \theta + c_{jk}.
\]

Equation 7 is Bock's (1972) nominal model, and it follows his terminology and that of Thissen and his colleagues (Thissen & Steinberg, 1984, 1986; Thissen, Steinberg, & Mooney, 1989). In the special case that is here called the two-parameter partial credit (2PPC) model,

\[
a_{jk} \equiv a_j(k - 1).
\]

The item discrimination, \( a_j \), can vary over items. Equation 9 implies that higher item scores reflect higher ability levels; that is, it is assumed that, for a given item, the higher the number assigned as an item score, the better the performance—an assumption consistent with the scoring rules used for the performance assessments. (In Bock's more general 1972 model, item scores are used as categories, and a higher item score does not necessarily mean better performance.) The following definition is also made:

\[
c_{jk} = -\sum_{i=1}^{k-1} y_{ji},
\]

where \( y_{ji} \equiv 0 \).

The 2PPC model can be seen to be either a special case of Bock's model or a generalization of Masters' (1982) partial credit model, which assumes that \( a_j = 1 \) for all \( j \). (Masters' assumption is analogous to the assumption made in the Rasch model used for multiple-choice items, which is that all items are equally discriminating.) The \( y_{ji} \) used here is equivalent to the \( \delta_j \) in Masters' notation. Thissen and Steinberg (1986) provide a comprehensive analysis of the relationships among Bock's, Masters', and others' models.

For the 2PPC model, there are \( m_j - 1 \) independent \( y_{ji} \) and one \( a_j \) for a total of \( m_j \) independent parameters estimated for each item. (The term two-parameter partial credit model is used to draw the analogy that, just as the two-parameter logistic model allows discriminations to vary over items but the one-parameter logistic model does not, the two-parameter partial credit model allows discriminations to vary over items while the partial credit model does not.) A marginal maximum likelihood procedure, implemented with the EM algorithm, was used to estimate the item parameters for the 2PPC items. (See Bock & Aitkin, 1981; Thissen, 1982, 1986.) Specifications for this procedure were written by the author and implemented in a microcomputer program, PARDUX, written by George Burket (1991). The accuracy of the program's estimates was validated by extensive simulation studies done in a manner analogous to those of Yen (1987) and by comparison to the estimates produced by MULTILOG (Thissen, 1986).

Maximum likelihood estimation based on the student's pattern of responses was used to produce trait estimates, \( \theta_n \). Because guessing was virtually nonexistent for these tests, the maximum likelihood trait estimate was a function of the weighted sum of the item responses, with the weights proportional to \( a_j \). (An analogous result for the maximum likelihood procedure occurs for the traditional two-parameter logistic model—e.g., Lord, 1980, pp. 76–77.)

Fit was evaluated with a statistic, comparing observed and predicted trace lines, which was a generalization of the \( Q_3 \) statistic described by Yen (1981). In addition, observed and predicted trace lines were compared graphically. The details of these fit analyses are not described here, but the 2PPC model was found to describe performance quite well for the vast majority of items. Items with poor fit were deleted from the scalings; there were two items so deleted from Form 41 and three items from Form 93.

**A Measure of LID**

The \( Q_3 \) statistic (Yen, 1984a) was used as a measure of LID; it is the correlation between performance on two items, after taking into account overall test performance. To calculate this statistic, a trait estimate, \( \theta_n \), is obtained for the \( n \)th student using the student's responses to all the items in the
scale. Using \( \hat{\theta}_a \) and the item parameters, the student’s expected performance on each item, \( E_{\mu} \), is determined, and the deviation, \( d_{\mu} \), between the student’s observed and expected item performance is calculated:

\[
E_{\mu} = E(X_j | \hat{\theta}_a) = \sum_{k=1}^{m_j} (k - 1) P_{jk}(\hat{\theta}_a), \quad (11)
\]

\[
d_{\mu} = x_{\mu} - E_{\mu}. \quad (12)
\]

The measure of LID for items \( j \) and \( j' \) is the correlation of these deviations taken over students,

\[
Q_{j\mu} = r(d_j, d_{j'}). \quad (13)
\]

It should be noted that including an item score both explicitly in \( x_{\mu} \) and implicitly in \( E_{\mu} \) (via the \( \hat{\theta}_a \)) produces artifactual negative \( Q_3 \) values. It can be demonstrated that the expected value of \( Q_3 \), when local independence is true, is approximately \(-1/(n - 1)\). Further information about the \( Q_3 \) statistic is presented in Yen (1984a).

**Results**

**Local Dependence of Items Within Passages: Reading**

The issue addressed here was whether the performance assessment items showed LID due to communality of passage and, if so, whether the magnitude of the LID differed from that seen with the traditional multiple-choice tests.

To explore this issue, for the six MSPAP Language Arts forms, the Reading items were sorted into passages, and the median and maximum within-passage \( Q_3 \) values were obtained. (For one \( n \)-item passage, there are \( n(n - 1)/2 \) \( Q_3 \) values.) The type of passage dependence was further classified as items related to a single passage, items related to two passages, and items related to a single passage in which students were Reading to Perform a Task. For the CTBS/4 RC test, items were classified by passages.

Results were similar across grades, and therefore pooled results are presented in Table 3 for medians and in Table 4 for maximum values. These results indicate there was little LID for the CTBS/4 tests. For the performance assessments, the \( Q_3 \) values for the single passages were quite similar to those for CTBS/4. The most LID tended to be for the items related to Perform-a-Task passages.

**Local Dependence of Items Within Passages: Mathematics**

Analyses within passages were conducted in the same way for Mathematics as they were for Reading. The results are given in Tables 5 and 6. There were very few CTBS/4 MC&A items linked to passages, but it is clear that the results for medians are quite similar for the performance assessments and CTBS/4. In terms of the maximum \( Q_3 \) values, there were proportionally far more LID items for the performance assessments than for CTBS/4. In Math Content, the very high \( Q_3 \) values were produced by items in which students were asked to do two very similar analyses of the same data set or in which they used the results of
TABLE 5

Stem and Leaf Plots of Within-Passage Median Q3 Values: Mathematics

<table>
<thead>
<tr>
<th>Performance Assessment</th>
<th>Math Content</th>
<th>Math Process</th>
<th>CTBS/4</th>
</tr>
</thead>
<tbody>
<tr>
<td>-.0</td>
<td>.0 01122</td>
<td>.0 01122235</td>
<td>-.0</td>
</tr>
<tr>
<td>.0</td>
<td>.0 000000001112222</td>
<td>.0 011111122333333</td>
<td>.0 001235</td>
</tr>
<tr>
<td>1</td>
<td>.1 037</td>
<td>.1 0</td>
<td></td>
</tr>
<tr>
<td>.2</td>
<td>.2 09</td>
<td>.2 8</td>
<td></td>
</tr>
</tbody>
</table>

Local Dependence of Follow-Up Math Process Items

A scaling was conducted for each of the six Math forms in which all Math Content and Process items were included. Items that measured Process only were identified and classified into two sets: (a) those Process items that required direct explanation of the preceding Math Content response and (b) those Process items that were not direct explanations of the preceding Math Content response. The Q3 value for each of these Process items with its preceding item was obtained. Table 7 presents the results of this analysis and shows that there was substantial local dependence between a Math Process item and its preceding item when direct explanation was involved.

TABLE 6

Stem and Leaf Plots of Within-Passage Maximum Q3 Values: Mathematics

<table>
<thead>
<tr>
<th>Performance Assessment</th>
<th>Math Content</th>
<th>Math Process</th>
<th>CTBS/4</th>
</tr>
</thead>
<tbody>
<tr>
<td>-.0</td>
<td>.0 0235567777899</td>
<td>.0 12222233577788</td>
<td>.0 001235</td>
</tr>
<tr>
<td>.0</td>
<td>.1 0133689</td>
<td>.1 0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>.2 12567</td>
<td>.2 8</td>
<td></td>
</tr>
<tr>
<td>.2</td>
<td>.3 00349</td>
<td>.3</td>
<td></td>
</tr>
<tr>
<td>.3</td>
<td>.4 28</td>
<td>.4</td>
<td></td>
</tr>
<tr>
<td>.4</td>
<td>.5 2</td>
<td>.5</td>
<td></td>
</tr>
<tr>
<td>.5</td>
<td>.6 19</td>
<td>.6</td>
<td></td>
</tr>
<tr>
<td>.6</td>
<td>.7 59</td>
<td>.7</td>
<td></td>
</tr>
<tr>
<td>.7</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Effects of LID on Test Information and SEM

Thissen, Steinberg, and Mooney (1989) and Sireci, Thissen, and Wainer (1991) have pointed out that an inappropriate assumption of local item independence will produce overestimates of test information and reliability and underestimates of the SEM. Thissen and his colleagues have examined the possible effects of LID on measurement error through the use of testlets. A testlet is a subset of the items in a test, and a testlet score is the sum of the item scores contributing to that testlet (Wainer & Kiely, 1987). When one or more testlets are formed, the items contributing to those testlets no longer appear as separate items in the total test score. Other items that do not contribute to testlets remain as separate items contributing to the total test score.

A testlet score is assumed to be independent of all other testlets and items. However, no assumptions are made about the independence of the items contributing to a testlet. Any LID that occurs among the items in a testlet is absorbed into that testlet score. If scaling is conducted, performance on a testlet is described with the same partial credit model used for separate items (e.g., Equation 7), with each score level being defined by one of the possible number-correct scores on that testlet.

If there is LID within a testlet, the testlet will not provide as much information as the sum of the information in the separate items. Thus, in the analyses conducted by Thissen et al. (Sireci, Thissen, & Wainer, 1991; Thissen, Steinberg, & Mooney, 1989) comparisons of measurement error were made between the test score in which the items were treated as independent units and that in which testlets were employed. If the testlet analysis produced more measurement error, then it was evidence that the assumption of local item independence was inappropriate and produced an underprediction of actual measurement error.

In the present study, analyses analogous to Thissen et al.’s (Sireci, Thissen, & Wainer, 1991; Thissen, Steinberg, & Mooney, 1989) were conducted. In addition, a further possible influence on measurement error was examined, based on the following reasoning. An analysis involving testlets can show less
test information than that involving separate items for two reasons. First, as pointed out by Thissen et al., the items actually are locally dependent so that the non-testlet analysis is overstating the information contributed by the items. Second, an analysis involving testlets may show greater SEM values than one involving separate items because important information is lost when the items are combined into the testlet.

The second reason merits further discussion. When item discriminations vary and/or there is guessing, there is more information in a score that considers the student's pattern of item responses than there is in one based on a number-correct score. For logistic models, this fact is described in theoretical terms by Lord (1980) and in practical terms by Yen (1984b) and Yen and Candell (1991). A similar result occurs for the 2PPC model. When items differ in terms of their discriminations, the trait estimate with maximum information will take into account the pattern of item responses. Thus, when items are combined into testlets, a greater SEM may occur because useful information in the pattern of item responses is lost. In the following analyses, this possibility was examined using a control scaling based on creating testlets using locally independent items.

Two tests (Reading Form 82 and Math Content Form 91) were chosen for these analyses. These tests had the greatest amount of LID. The Reading test had 31 items, 13 of which were identified via the Q3 statistic as being locally dependent. These items were placed in testlets. There were five testlets: three involving item pairs, one involving a triplet, and another involving a quadruplet; the average Q3 value for the items entering into these testlets was .24. The Math Content test had 52 items, 19 of which were placed in testlets. There were eight testlets: five involving item pairs and three involving triplets; the average Q3 value of the items entering into these testlets was .29.

Each of these tests was scaled three times. The first scaling (Items) treated all the items as separate (putative independent) units. The second scaling (LID Testlets) used the testlets described in the preceding paragraph along with the remaining items that did not appear in testlets. The third scaling (Non-LID Testlets) involved as many testlets as appeared in the LID Testlet analysis, but the items placed in these testlets were selected to be as locally independent as possible. The mean Q3 value of the items in these testlets was −.03. The items used in these testlets were matched as closely as possible to those in the LID Testlet analysis in terms of number of levels, difficulty, level of discrimination, and variation in discrimination. The within-testlet variability in item discrimination (square root of the pooled within-testlet variance) for Reading Form 82 was .15 for the LID Testlets and .17 for the Non-LID Testlets, and for Math Form 91 it was .25 for the LID Testlets and .33 for the Non-LID Testlets.

Within each content area, these three analyses were based on the same data set. The results of these three scalings were aligned through the marginal maximum likelihood estimation procedure, which employs a theoretical distribution of true thetas with a mean of 0 and a standard deviation of 1. After the scaling, all three analyses were linearly transformed with the same constants so that estimated trait values would have an approximate mean of 500 and standard deviation of 50.

The Items analysis was compared with each of the testlet analyses in terms of relative efficiency as described in Tables 8 and 9. To obtain the relative efficiency, the area under the information function for each item or testlet was obtained, and then these areas were summed for the appropriate sets. For the Testlets' relative efficiency, the sum taken over all testlets in the LID Testlets or Non-LID Testlets analysis was compared with (divided by) the sum taken over the items in the Items analysis that corresponded to the items assigned to those testlets. The sums for the Non-testlets' relative efficiency were based on the items not included in the Testlets' relative efficiency. The total relative efficiency was based on sums taken over all items and testlets in a scaling. Mean item discriminations were compared in a similar manner. The ratios reported are less than 1.0 if the Items’ scaling produced more putative information than the scalings involving testlets.

The results in Tables 8 and 9 are very clear. For the items in the testlets, the putative information provided by the scaling of the separate items (Items analysis) is much greater than the information provided by the LID Testlets analysis, with relative efficiencies of .61 (Reading) and .66 (Math Content). The results for the Non-LID Testlets analysis indicate that, if testlets are formed from independent items, there is little effect on the information (i.e., relative efficiencies of .98 and .96); that is, it is not the formation of testlets in and of itself that produces a reduction in information.

A pair of items from the Math Content Form 91 test was identified that had an extreme amount of LID, Q3 = .57. That pair of items formed one of the testlets in the LID Testlets scaling, and it had scores ranging from 0 to 3. Figure 1 displays the information function of this testlet, scaled the two ways. The information from the LID Testlets analysis is half the sum of the information from the separately scaled items.

### TABLE 8
Comparisons of Information and Discrimination for Scalings with Testlets with a Scaling Without Testlets: Reading

<table>
<thead>
<tr>
<th>Scaling</th>
<th>LID Testlets</th>
<th>Non-LID Testlets</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Testlets</td>
<td>1.03</td>
<td>1.06</td>
</tr>
<tr>
<td>Non-testlets</td>
<td>1.01</td>
<td>1.00</td>
</tr>
<tr>
<td>Relative Efficiency*</td>
<td>.61</td>
<td>.98</td>
</tr>
<tr>
<td>Testlets</td>
<td>1.03</td>
<td>1.06</td>
</tr>
<tr>
<td>Non-testlets</td>
<td>1.01</td>
<td>1.00</td>
</tr>
<tr>
<td>Total</td>
<td>.87</td>
<td>1.03</td>
</tr>
<tr>
<td>Ratio of Mean Item Discrimination*</td>
<td>.63</td>
<td>.96</td>
</tr>
<tr>
<td>Testlets</td>
<td>1.01</td>
<td>1.00</td>
</tr>
<tr>
<td>Non-testlets</td>
<td>1.01</td>
<td>1.01</td>
</tr>
</tbody>
</table>

*Compared with Items scaling that had no testlets.
TABLE 9
Comparisons of Information and Discrimination for Scalings with Testlets with a Scaling Without Testlets: Mathematics Content

<table>
<thead>
<tr>
<th>Scaling</th>
<th>LID Testlets</th>
<th>Non-LID Testlets</th>
</tr>
</thead>
<tbody>
<tr>
<td>n Testlets</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>n Non-testlets</td>
<td>33</td>
<td>33</td>
</tr>
<tr>
<td>Relative Efficiency*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Testlets</td>
<td>.66</td>
<td>.96</td>
</tr>
<tr>
<td>Non-testlets</td>
<td>1.07</td>
<td>1.00</td>
</tr>
<tr>
<td>Total</td>
<td>.90</td>
<td>.98</td>
</tr>
<tr>
<td>Ratio of Mean Item Discrimination*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Testlets</td>
<td>.64</td>
<td>.95</td>
</tr>
<tr>
<td>Non-testlets</td>
<td>1.05</td>
<td>1.00</td>
</tr>
<tr>
<td>Total</td>
<td>.91</td>
<td>.95</td>
</tr>
</tbody>
</table>

*Compared with Items scaling that had no testlets.

It should be kept in mind that these tests and items were selected for these analyses because they showed the greatest amount of LID. Most tests would show much less effect from LID than those examined here.

The implications of these results vary according to the use that is made of the predicted information or SEM values. Because test information is roughly proportional to numbers of items, it is sometimes used to predict the change in test length needed to reach a desired level of score accuracy. It is clear that predictions made for locally dependent items can greatly underestimate the necessary test length.

On the other hand, if SEM values are being used as descriptive statistics to provide general guidance about score accuracy, the seriousness of ignored LID can be less severe. This effect can occur because only some of the items in the test are locally dependent. For example, consider the Reading test that was the basis of the analyses in Table 8. This test had more items involved in testlets than any of the other reading tests; of its 31 items, 13 were in one of its five testlets. Figure 2 displays the SEM functions for that test for both the LID Testlets and Items scalings and shows that these SEM functions are quite similar. In many applications, the misstatement of a SEM by that amount would have minor practical effect.

Effects of LID on Item Discriminations

Tables 8 and 9 in the preceding section showed that testlets based on LID items tended to have substantially lower discriminations than the separately scaled items. In the example in Figure 1, the testlet discrimination was approximately 50% of the mean of the discriminations of the items contributing to that testlet. It should be noted that this effect was among the most extreme seen in any of the tests.

In general, the Perform-a-Task items showed the greatest amount of LID among the Reading items. The discriminations of those items were examined to determine whether their LID was contributing to unusually high discriminations. For each of the six forms in this study, the mean Q and the mean discrimination (a) was found for each Perform-a-Task passage. The mean discrimination for items not related to the Perform-a-Task passages was also obtained; these means were similar across forms, with a grand mean of 1.11. The Perform-a-Task values (mean Q, mean a) were the following for the six passages: (.05, 1.23), (.06, 1.12), (.20, 1.63), (.23, 1.12), (.25, .98), (.29, 1.71). There was no systematic trend for the Perform-a-Task passages to have unusually high discriminations or for these discriminations to be related to the amount of LID in the passage.
Effect of LID on Trace Lines

The probability of a student's receiving any given total raw score for the two items combined into a testlet, \( Y = X_1 + X_2 \), can be predicted, assuming local item independence, to be

\[
P(Y = y|\theta) = \sum_{x_1=0}^{m_1-1} P(X_1 = x_1|\theta) P(X_2 = y - x_1|\theta).
\]  

(14)

The prediction based on Equation 14 can be compared with empirical testlet results to determine if LID is affecting trace lines.

The accuracy of Equation 14 was examined for a testlet in which the items were highly locally dependent (i.e., the pair of items from Math Content Form 91 that appear in Figure 1). Figure 3 displays the trace lines for each of the scores for that testlet based on the LID Testlets analysis. For the sake of clarity, separate figures display the separate score levels in Figure 4. Also appearing in Figure 4 is, for each trace line, the prediction based on the Items scaling and Equation 14, which assume local item independence. It can be readily seen that, in this example where there was extreme LID, the assumption of local independence produced inaccurate predictions, particularly for scores of 2 and 3.

Effect of LID on Characteristic Functions

Figure 5 displays the characteristic function for the Math Content Form 91 testlet described in the preceding section. One version of this function is based on the Items scaling and Equation 6. The other characteristic function is based on the LID Testlets scaling. In contrast to the great effect of LID on the accuracy of the trace line predictions, there is little effect on characteristic functions. This conclusion was supported by the examination of test characteristic functions as well; LID has little effect on the relationship predicted between raw scores for tests and trait values.

Study Limitations

These analyses are based on one set of performance assessment and multiple-choice achievement data. To determine the generality of these results, it would be informative to conduct similar analyses on other assessment systems. While some further information about the performance of Q3 with simulated and real data is available (Yen, 1984a), that information is based on multiple-choice
Managing Local Item Dependence

The following six procedures are possible ways of reducing LID or, when that is not appropriate or feasible, of analyzing data so that the LID has minimal negative effects on measurement characteristics. All these procedures have limitations, which are discussed. The procedures are ordered in terms of when each would occur during the test construction or analysis process. The advantage of managing LID later in the testing process is that the procedure has less effect on the test design, administration, or scoring rule definition or implementation. The major disadvantages of managing LID after data are collected are inefficiency and inadequate data for meeting a particular goal; for example, time and money may have been invested in the collection and scoring of separate items when these separate data prove not to be very informative.

(1) Construct Independent Items

As discussed earlier, much of the philosophy of performance assessments is opposed to the idea of constructing independent items. However, if achieving the goals of the assessment requires independent items, steps can be taken to move the assessment in that direction.

From the perspective of the generalizability of scores to a variety of real life situations, it is ideal to have multiple independent tasks (or passages) within a test form, with the tasks varying along relevant dimensions in terms of the context in which they measure the performance of interest. However, establishing the setting for each task takes a substantial amount of time and necessarily limits the number of such tasks a student can take. It is often possible to create multiple independent items within a task and still be consistent with principles of authenticity. Particularly for younger students, many real life tasks involve multiple smaller steps or activities that can be realistically reflected in the assessment.

It is clear that the multiple-choice tests described here had items that displayed little LID due to passages. For the MSPAP Reading performance assessments, efforts were taken to write the items so that they provided as much
independent information as possible, and, for the most part, these efforts were successful, showing little LID among the reading items. The outstanding exception to the finding of little LID was the items relating to Performing a Task, which were much more locally dependent than others. Reading to perform a Task appeared to involve student skills beyond those involved in Reading to Inform or Reading for Literary Experience.

In Mathematics performance assessments, there was a great deal of LID. In constructing these tests, tasks were made independent, but, in order to achieve greater authenticity, the items within tasks were not necessarily constructed to be independent. Items that involved direct explanation of previous items showed substantial LID, as did those that asked students to perform very similar types of analyses with the same data set.

If multiple items are created, but these items are not independent, little may be gained and much may be lost in terms of the validity of the student performance, student testing time spent on multiple responses, rater time evaluating these responses, and psychometrician time spent handling the multiple responses. The maximum benefit in terms of efficiency of testing is to use only those items that are needed and have them be independent.

(2) Administer Test Under Appropriate Conditions

It is obvious that to avoid LID due to speededness, fatigue, or undesired interference or assistance appropriate test administration conditions are needed. In some instances, peer interaction can be a legitimate part of the assessment, as when the results are used to evaluate groups, such as schools, rather than individual students. The LID that results from such interactions would require special analysis procedures. It would also be the case that the amount of information provided by such dependent responses would not be as much as that produced by the same number of independent observations (students × items).

(3) Combine the Grading of LID Items

Efficient grading of students' constructed responses must be considered because it is a very expensive part of performance assessments. Consider the case of the Math items that involved direct explanation of preceding items and that provided limited independent information. While asking students for such explanations can be a very important contribution to the validity of the test, it may be that the grading of such explanations should be combined with the grading of the preceding item. In other words, it may be more efficient to combine such items at the time they are scored by raters than at the time the scales are produced.

There are limitations to the idea of combining the grading of LID items. In some cases, it may be difficult to determine a priori which responses are locally dependent. Also, it can take additional rater time and decrease the quality of the ratings if a scoring rubric becomes too complicated.

(4) Review Tests to Identify LID Items

As part of the MSPAP program, scales were produced for Reading and Mathematics content areas across 17 test forms. In preparation for these scalings, test forms were scrutinized to identify a priori items that appeared likely to display LID. This process was difficult and time consuming. When data (i.e., Q3 values) were examined, the proportion of these items that actually displayed LID was much smaller than anticipated, particularly in Reading. The within-passage results reported here display that effect. Thus, reviewing the items empirically using the Q3 values was much more efficient than reviewing them a priori.

(5) Construct Separate Scales

In one case, the a priori identification of items likely to be LID was very effective. In Mathematics, there were a substantial number of items that required students to explain the reasoning used to obtain a preceding answer; these were all Math Process items. Because it was believed likely that these items would show high LID on the Math Content items, the items were separated into analyses for two scales, Math Process and Math Content. As described in an earlier section, the direct explanation items showed substantial empirical LID on their preceding Math Content items, justifying the construction of the two scales.

An alternate strategy would have been to construct testlets for all these items. The disadvantage of the testlet strategy would have been related to outcome scores. It was desired to have Math Content and Process outcome scores that had as much discriminant validity as possible. Constructing the separate scales did not remove the implicit dependence of the Process items on the Content items. However, if the Content and Process items had been combined into testlets, these testlets would then be applied to both Content and Process outcomes. Thus, using testlets, the Math Process performance would have been included in the Math Content outcomes, and more Math Content performance would have been included in the Math Process outcomes than was necessary. The use of separate scales permitted clearer distinctions to be made between the Content and Process outcomes than would have been possible with the use of testlets.

The use of separate scales did not eliminate the need to consider LID. To obtain a Math Total score for reporting purposes, the Math Content and Math Process scores were averaged. In producing SEM values for this average, it was necessary to consider the fact that measurement error for the two scales was correlated.

(6) Use Testlets

The rule of thumb that was used in the operational MSPAP scalings was that testlets would be created from item sets that would normally be administered together and that had Q3 values ≥.20. (In some instances, the model fit was significantly worse for the testlet than for the separately scaled items; in such cases, the scaling of the separate items was retained.) Across the 17 Reading forms, 13% of the items were placed in testlets. Across the 15 Math forms, 25%
of the Math Content items and 12% of the Math Process items were scaled as testlets.

A testlet does not remove the LID among the items in the testlet; it merely provides a more accurate way of relating the performance on that set of items to the other items in the test. If the items in a testlet measure different outcomes or objectives, then the testlet score must be applied to all the outcomes it measures. Such an increase in the breadth of the measures contributing to an outcome score can diminish the clarity of the meaning of the outcome score.

The loss of information by the use of testlets can be minimized if only those items that actually are locally dependent are included in each testlet. That is, if only some of the items related to a passage show LID, all the items related to that passage should not be combined into a testlet; only the LID items should be combined. Many "small" testlets, each involving few items, are likely to retain more information than one "large" testlet.

Testlets produced higher SEM values than did the scaling of separate items. Based on these analyses and those of Thissen et al. (Sireci, Thissen, & Wainer, 1991; Thissen, Steinberg, & Mooney, 1989) it appears likely that the higher SEM value is more reflective of reality. In the worst case, for the LID items in the testlets, the effect was dramatic, with the individual item scaling overstating the information by 52 to 64%. The fact that LID can produce such inefficiency motivates the suggestion for constructing independent items. While the in-


Author

WENDY M. YEN is Director, Research and Measurement Services, CTB Macmillan/McGraw-Hill, 20 Ryan Ranch Road, Monterey, CA 93940. Degrees: AB, MA, PhD, University of California, Berkeley. Specializations: item response theory and test theory.