# DOCUMENT RESUME

| | |
|---|---|
| ED 453 236 | TM 032 587 |
| AUTHOR | Allen, Sally; Sudweeks, Richard R. |
| TITLE | Identifying and Managing Local Item Dependence in Context-Dependent Item Sets. |
| PUB DATE | 2001-04-13 |
| NOTE | 31p.; Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001). |
| PUB TYPE | Reports - Evaluative (142) -- Speeches/Meeting Papers (150) |
| EDRS PRICE | MF01/PC02 Plus Postage. |
| DESCRIPTORS | *College Students; Error of Measurement; Higher Education; Physics; Reliability; *Scoring; *Test Items |
| IDENTIFIERS | *Item Dependence |

ABSTRACT

A study was conducted to identify local item dependence (LID) in the context-dependent item sets used in an examination prepared for use in an introductory university physics class and to assess the effects of LID on estimates of the reliability and standard error of measurement. Test scores were obtained for 487 students in the physics class. The test consisted of 24 context-dependent sets containing 3, 4, or 5 multiple choice items linked to a unique information display to yield 100 multiple choice items. The method used to identify LID was based on inter-item correlation matrices, with a matrix of the mean inter-item correlations constructed for each of five ability levels. Study results provide evidence that is possible to write context-dependent items that do not contain LID. Only 10 of the 24 item sets exhibited evidence of LID. If items do contain LID, it is possible to score the test in a way that will reduce its effects. Scoring the item sets as individual items or as testlets is not the best solution, but the benefits of using testlets can be gained by using mixed scoring gained from using the testlets accompanied by a procedure for estimating reliability. (Contains 22 references.) (SLD)

# IDENTIFYING AND MANAGING LOCAL ITEM DEPENDENCE IN CONTEXT-DEPENDENT ITEM SETS

by

Sally Allen,

Kennewick School District,

Kennewick, WA

and

Richard R Sudweeks,

Brigham Young University,

Provo, UT

Paper presented at the annual meeting of the

American Educational Research Association,

Seattle, WA

April 13, 2001

# IDENTIFYING AND MANAGING LOCAL ITEM DEPENDENCE IN CONTEXT-DEPENDENT ITEM SETS

Context-dependent sets of test items proved a versatile means of assessing examinees' ability to apply their knowledge including their skill sin analytical thinking and problem solving. A context-dependent item set consists of two parts: (a) an information display, and (b) a series of test items that require the examinees to mentally process the information in the display (Halyadyna, 1992; Wesman 1971). The information display provides the contest or setting within which the problems presented in the test items are situated. Together, the items and the accompanying information provide a problem situation for the examinees to think or reason about to correctly answer each question (Haladyna, 1992; Sireci, Thissen, & Wainer, 1991).

Examples of two context-dependent item sets are displayed in Appendix A. Some scholars have used labels such "item bundles" (Wilson & Adams, 1995), "interpretive exercises" (Ebel, 1951; Wesman 1971), "testlets" (Lee & Frisbie, 1999), "passages" (Yen, 1993), and "item clusters" (Ferrara, Huynh, & Baghi, 1997) to sets of test items that are dependent upon a common information base.

Together, the items and the accompanying information provide a situation in which the students are expected to think or reason in order to answer each question (Sireci, Thissen, & Wainer, 1991; Haladyna, 1992; Worthen, White, Fan, & Sudweeks, 1999).

Due to the fact that all the items in an item set are dependent on the same context, these items are likely to be locally dependent on each other. Local item dependence (LID) occurs when an examinee's performance on one item influences their performance on subsequent items in the test (Thissen & Wainer, 1996; Yen, 1993). One method used to manage scores from a test containing context-dependent item sets is to use testlet scoring. This procedure involves scoring each set of items as if it were a single item (Wainer & Kiely, 1987).

Statement of Purpose

The purpose of this study was twofold. First, to identify LID in the context-dependent item sets used in an exam prepared for use in an introductory, university physics class. Second, to assess the effects of LID on estimates of the reliability and standard error of measurement. To manage LID it is necessary to first identify which specific item sets exhibit evidence of LID and to determine to what degree it is present. An accurate estimate of reliability is needed when using test scores to make valid, dependable decisions. If a scoring method can be found that results in an accurate reliability estimate, the negative effects of LID can be minimized.

Research Questions. This study focused on answering the following questions:

1. To what extent do the context-dependent item sets used in the Physics 105 final examination possess LID?

A. How many of the 24 item sets in the test show evidence of LID?

B. What is the average amount of LID in the 24 item sets, and how does this amount vary from set to set?

2. To what extent does the presence of LID lead to an overestimate of the reliability of the scores obtained from the Physics 105 examination as evidenced by the difference in the magnitude of Cronbach's alpha for item-based scoring (item $^{s}x$) and the generalizability coefficient?

3. How does the use of testlet-based scoring affect the estimated reliability of scores obtained from the Physics 105 examination?

A. If testlet-based scoring is used for all of the item sets including both the sets that show evidence of LID and those that do not posses LID, how closely does the Cronbach's alpha for testlet-based scoring (testlet $^{s}x$) match with the generalizability coefficient?

B. If testlet-based scoring is used only for item sets that show evidence of LID and item-based scoring is used for item sets that do not contain LID, how closely does Cronbach's alpha for mixed scoring (mixed $^{s}x$) match the generalizability coefficient?

## Literature Review

### Local Item Dependence

Both classical test theory and modern item response theory are based on the assumption that the items in a test are locally independent (Crocker & Algina, 1986; Ferrara, Huynh, & Michaels, 1999; Yen, 1993). However, the use of context-dependent items may violate this assumption due to the fact that all the items within a given set are linked to the same scenario or problem situation, these items are likely to be locally dependent. The common context may introduce extra dependence to those items sharing it. That extra dependence implies that the items may be measuring some extraneous construct in addition to the construct that the test as a whole is intended to measure (Wainer & Thissen, 1996). LID occurs when an examinee's performance on an item influences their performance on subsequent items in the test.

For the assumption of item independence to be true, an examinee's performance on one item must not be affected by their responses to any other items on the test. Performance on a given test item should only be influenced by the examinee's ability and the characteristics of that item (Hambleton, 1989).

It is important to note that the concept of local independence is defined in terms of individual examinees or examinees at a given ability level. The assumption of LID does not imply that a set of test items are uncorrelated over the total group of examinees. What this assumption requires is that the item responses be uncorrelated for examinees at any given ability level (Hambleton, 1989).

Modern testing is moving towards the idea that larger, more holistic tasks such as context-dependent item sets should be the fundamental units of which tests are made (Sireci, Thissen, & Wainer, 1991). When using item sets, educators should be aware of the risks of LID.

Undesirable Effects of LID

Sireci, Thissen, & Wainer (1991) showed that failure to account for LID leads to overestimation of the reliability of the scores obtained from a test. They claimed that the overestimation can be as high as 10-15%. Yen (1993) also showed that the presence of LID in a test can affect the statistical properties of the scores.

In classical test theory reliability is defined as the degree of consistency in two or more measures of the same trait obtained from the same individuals. One way to estimate reliability is to use Cronbach's alpha, which provides an estimate of inter-item consistency. If the item-based method of computing the coefficient alpha is used, then the reliability of scores from a test composed of locally dependent items will be overestimated. If an item, at a given ability level, has a high correlation with the total score and is also highly correlated with another item, then that second item will also be highly correlated with the total score. An overestimate of the reliability may lead to the misinterpretation of an examinee's scores. The scores will likely be viewed as being more consistent and dependable than they actually are (Lee & Frisbie, 1999).

Classical test theory is based on the idea that an examinee's observed score on a test is a composite consisting of the sum of that person's true score and an error score. The formula is:

$$X_i = T_i + ,_i , \tag{1}$$

where $X_i$ represents that examinee's observed score, $T_i$ represents an examinee's true score, and

$,_i$ represents that examinee's error score. This assumption specifies that the error scores for a group of examinees are uncorrelated with the examinees' true scores. This assumption also specifies that error scores from two separate tests are uncorrelated with each other. However, since error scores are not directly observable, it is impossible to test this assumption directly. More often than not, this assumption is violated and the errors are inter-correlated (Zimmerman & Williams, 1980). The concern is with the size of the correlations and how the magnitude of this correlation affects the reliability of the measurements. Relatively low correlations between the errors do not greatly reduce reliability. However, moderate or large correlations do reduce reliability and therefore the scores are less dependable.

The standard error of measurement (SEM) is an estimate of intra-individual variability. If an examinee were tested repeatedly, their obtained score would likely vary from one testing occasion to another. The SEM is defined as the square root of the average squared difference between a student's true score and their observed scores on a test. It can be estimated by using the standard deviation and reliability coefficient of the scores obtained from a group of examinees. If LID is present the actual amount of measurement error in the total scores will be greater than what it is calculated to be by using the classical SEM. The SEM is often used to construct confidence intervals as a basis for interpreting the scores of individual examinees. If the error scores for a group of examinees are correlated with their true scores then LID is present and this undesirable situation will likely cause users of the test scores to overestimate the precision of each examinees' score (Yen, 1993).

### Managing LID

There are several ways to reduce the amount of LID in test items and the effect it has on the statistical properties of the test scores. The most obvious solution is to write independent items. However, writing good independent items is a difficult task, particularly in context-dependent item sets, since the items are designed to all be related to a common information base. Rewriting questions that have been identified as having LID is also a good idea. However, some types of items may be interdependent by their very nature. Usually a test is administered and the scores collected before LID is identified, therefore it is not feasible to go back and re-administer the test with rewritten questions. Furthermore, there may be times when test writers deliberately include dependent items in an effort to construct an authentic assessment.

Testlet-based Scoring. One method used to analyze scores from a test containing context-dependent item sets is to use testlet scoring. This procedure involves scoring the items within each set as if they were a single item. The examinee's responses to the *m* items in a set are assigned a single score between 0 and *m* (Wainer & Kiely, 1987). In testlet scoring each item set is assumed to be independent of other item sets in the test as well as independent of individual items in other parts of the test. Testlet scoring does not remove the LID among the items within an item set. The LID is absorbed by the testlet score and it merely provides a more accurate way of relating the performance on that set of items to the other item sets and items on the test (Yen, 1993)

However, there are disadvantages to testlet scoring. Yen (1993) demonstrated that testlet scoring reduces estimates of the reliability of the composite scores. Andrews (1992) showed that the use of testlets rather then individual item scores underestimates the reliability of the scores, if the item sets contain LID. Testlet scoring has the equivalent result of reducing the effective length of the test (Guilford, 1936). The length of the test is determined by the number of testlets, rather than the number of individual items. Generally, the longer a test is, the more reliable it is. When there are more items, the individual items (or testlets) have a higher correlation with the true scores, thus having a higher reliability (Nunnally, 1967).

Another disadvantage of testlet scoring involves item information. Item sets scored as a single item tend to have a substantially lower discrimination than the individually scored items. When an item set is scored as if it were a single item, the response of the student to each individual item within the set and the exact information contributed by each item is lost. Only information on the student's responses to the item set as a whole is provided; details on what the student does and does not know and understand are not provided (Yen, 1993).

However, the lost information and reliability reduction can be minimized if only those item sets that include LID are scored as testlets. Yen (1993) showed that if only small item sets that contained LID were scored as testlets and then combined with individual independent items then the scores would be apt to retain more item information, produce a longer test, and hence give a more reliable score.

Assessing the Effects of LID

The effects of LID on reliability can be assessed by computing an accurate estimate reliability by using generalizability theory (G-theory). G-theory provides a methodology that disentangles multiple sources of error in measurement and shows the degree of influence of each. G-theory differentiates and decomposes the various sources that contribute to the total observed variance into independent parts called variance components. G-theory computes a separate variance component for each random effect. The process differentiates between the variance associated with the different examinees and the error contributed by the item sets. Cronbach's alpha does not consider the item set effect as a separate source of variation. It ignores the item sets as a separate effect, and therefore ignores the variation they contribute. Hence the item set variation is included in the person variation and inflates the estimated reliability coefficient. G-theory accounts for the variance from item sets and therefore produces a more accurate reliability estimate that was used as a standard in this study. The variance components are used to calculate error variances which provide the basis for computing reliability-like coefficients (Lee & Frisbie, 1999).

The relative error variance produced in a generalizability theory study is analogous to the error variance of classical test theory (Brennan, 1992 & Shavelson & Webb, 1991). In G-theory the reliability estimate based on relative error variance is referred to as the generalizability coefficient (G-coefficient). G-theory also defines an estimate of reliability for absolute decisions, but it is not used in this study.

G-theory partitions the global measurement error into variance components represented by the different facets and interactions included in the design of a study. Since the G-coefficient is not inflated by LID it provides a more accurate estimate of the reliability of the scores than Cronbach's alpha coefficient. Comparison of the G-coefficient and Cronbach's alpha for the individual items on a test indicates to what degree a classical reliability estimate computed from individual item scores overestimates the reliability of the scores.

An alpha coefficient can also be used to compute a reliability estimate for scores resulting from testlet-based scoring. Comparison of testlet reliability and a G-coefficient will show how much a Cronbach's alpha coefficient computed from testlet scores underestimates the reliability of the scores (Lee & Frisbie, 1999).

## Method

### Sample

Test scores were obtained for 487 students enrolled in Physics 105 at Brigham Young University during Winter semester 1999 and used for the analysis in this study.

### Instrumentation

The test used was prepared as a final, comprehensive examination intended to assess the whole range of topics and objectives taught in Physics 105. The test consisted of 24 context-dependent sets each containing 3, 4, or 5 multiple choice items linked to a unique information display. All together, the 24 items sets containing 100 multiple choice items. The displays consisted of a written scenario, a pictorial display, or a combination of both. Explicit instructions were given for each context-dependent item set stating how to use the information for that item set and which items were linked to that information (A specimen copy of one item set from the test is shown in Appendix A).

### Analysis

The analysis was conducted in two stages. The first stage focused on determining to what extent LID was present within each of the 24 item sets. The second stage examined the consequences of LID on reliability estimates and how testlets can be used to manage any negative effects.

Identifying LID. The method used to identify LID was based upon inter-item correlation matrices. All of the items were grouped into sets according to the context in which they were presented. The 487 students were classified into five, ordered ability levels based on their total score on the test. The ability levels were created in such a way as to have about the same number of students in each level. To control for variation due to heterogeneous groups, scores that were too extreme were deleted from the sample. Thirty-five students with scores between 23 and 45 as well as six students with scores between 89 and 93 were deleted. Then five ability groups were defined. The score range and the resulting number of students in each ability level are shown in Table 1.

Table 1.
Number and Percent of Students by Ability Level

| Ability Level | Score Range | Number of Students | Percent of Students |
|---|---|---|---|
| 1 | 46 – 56 | 84 | 18.8 |
| 2 | 57 –64 | 95 | 21.3 |
| 3 | 65 – 71 | 94 | 21.0 |
| 4 | 72 – 78 | 91 | 20.4 |
| 5 | 79 – 88 | 83 | 18.6 |
| Total | | 447 | 100 |

A matrix of the mean inter-item correlations for each ability level was constructed. The first step in computing mean inter-item correlations was to create a 100 by 100 matrix for each ability level displaying the correlation coefficients for all pairs of items. Each matrix consisted of the inter-item correlations for all 100 items. Refer to the top matrix of Figure 1 for an example of that matrix. This simplified diagram displays a piece of the whole 100 by 100 matrix for one ability level. Two item sets, consisting of five items and four items respectively, are included in the example shown in Figure 1. The scores for each of the five items in the first set were correlated with the scores for each of the four items in the second set. The resulting between-set correlations are shown in the unshaded sub-matrix in the lower left quadrant of the top matrix. The scores for each of the five items in the first set were also correlated with each other. The shaded sub-matrix on the upper left displays these within-set correlations for the five items in the first set. Similarly, the shaded sub-matrix in the lower right portion of the top matrix displays the within-set correlations for the four items in the second set. For the sake of simplicity

| Items | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | | | | | | | | |
| 2 | 0.01 | 1 | | | | | | | |
| 3 | 0.49 | 0.19 | 1 | | | | | | |
| 4 | 0.11 | 0.17 | 0.32 | 1 | | | | | |
| 5 | 0.06 | 0.17 | 0.13 | 0.20 | 1 | | | | |
| 6 | -0.07 | -0.02 | -0.04 | 0.24 | 0.02 | 1 | | | |
| 7 | -0.08 | -0.13 | -0.04 | -0.12 | -0.10 | -0.03 | 1 | | |
| 8 | -0.11 | -0.02 | -0.05 | -0.07 | -0.11 | -0.09 | 0.68 | 1 | |
| 9 | -0.06 | 0.07 | -0.03 | -0.08 | -0.11 | -0.09 | 0.55 | 0.41 | 1 |

| Items | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1–5 | Average Inter-item Correlation Within Set 1 = .185 | | | | | | | | |
| 6–9 | Average Inter-item Correlation Between Sets = -.049 | | | | | Average Inter-item Correlation Within Set 2 = .240 | | | |

Figure 1. Example of Inter-item Correlations for Two Item Sets in One Ability Level

only the non-redundant correlations are displayed (see the shaded sections of the top matrix of Figure 1). The non-redundant inter-item correlations in each of the sub-matrices were averaged. Compare the shaded portions of the top and bottom diagrams in Figure 1 as an example. The 10 correlation coefficients were summed and then divided by 10 with a result of .185 (as displayed in the bottom matrix).

Compare the unshaded portions of the top and bottom matrices of Figure 1 as an example. The average inter-item correlation between sets 1 and 2 was calculated using all 20 inter-item correlations between sets 1 and 2 (see the top matrix). The resulting coefficients were then summed and divided by 20 with a result of -.049 (see the bottom matrix). A similar procedure was followed to compute the average between and with-in set correlations for all possible sets.

It is important to note that some correlations were incalculable. When all students give the same answer for a specific question, then the item score has no variance and a correlation can not be computed for that particular item, because by definition such an item does not correlate or co-vary with other items. This occurred whenever an item was so easy that all students at a particular ability level answered it correctly and when an item was so difficult that no one in a given ability group answered it correctly. Any incalculable correlations were left out of both the numerator and the denominator in computing the average within-set and between-set correlations.

After a correlation matrix was computed at each ability level, the average within-set correlations for the various ability levels were summed and divided by five, the number of ability levels. The same procedure was used to compute the average between-set correlation (Ferrara, Huynh, & Baghi, 1997 and Ferrara, Huynh, & Michaels, 1999).

Assessing the Effects of LID and the Advantages of Using Testlet Scoring. To assess the effects of LID on reliability and the advantages of using testlets, generalizability theory (G-theory) was applied. The data for this analysis can be described by a two-facet, unbalanced p x (i:s) design. In this notation, the letter *p* represents *persons*, or students. Since they are the object of measurement, they are not considered to be a source of measurement error. The letters *i* and *s* represent *items* and *item sets* respectively. They are both considered to be random facets. A variance component for persons and item sets was computed. However, since items were nested within item sets, it was not possible to compute a separate estimate of the variance component for items that was free from other sources of variance. The best the researcher could do was to compute a single variance component for *items within sets* that confounded these two sources of error.

The design was unbalanced due to the fact that the number of items varied across item sets. The urGENOVA software developed by Brennen (1999) was used to estimate variance components for all of the random effects. The variance components were used to estimate a relative error variance. Because of the unbalanced design the following equation given by Lee and Frisbie (1999) was used to compute an estimate of the relative error variance:

$$ {}^{S}s^{2}\ ({}^{L}{}_{F}) = 1/\,n_{+}\ \ [V_{i} \text{ x } {}^{S}s^{2}_{(ps)} + {}^{S}s^{2}_{(pi:s)}] \ \ \text{where } V_{i} = {}^{N}{}_{L}n_{i:s}^{2}/\,n_{+}\,. \qquad (2) $$

The symbol $n_{+}$ indicates the number of items in the test, ${}^{S}s^{2}_{(ps)}$ represents the variance component for persons crossed with item sets, and ${}^{S}s^{2}_{(pi:s)}$ represents the variance component for persons crossed with items, the items being nested in item sets.

Classical test theory reliabilities were also estimated. Cronbach's alpha for item-based scoring (item ${}^{s}x$) was computed. Item ${}^{s}x$ is a reliability coefficient computed from the 100 individual items in the exam.

Then Cronbach's alpha for testlet-based scoring (testlet ${}^{S}x$) was computed. To compute testlest ${}^{S}x$, the sum of the variances for the 24 testlets was substituted in the reliability formula in place of the sum of the item variances.

Finally, testlet and item-based scoring were combined. A mixed ${}^{s}x$ reliability coefficient was computed using a mixture of the testlet and item-based scoring procedures. Testlet-based scoring was applied to those 10 item sets identified as possessing LID. The remaining items were scored separately as individual, independent items (see Figure 2.)

| Item Sets | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.09 | | | | | | | | | | | | | | | | | | | | | | | |
| 2 | 0.00 | 0.21 | | | | | | | | | | | | | | | | | | | | | | |
| 3 | 0.02 | 0.04 | -0.01 | | | | | | | | | | | | | | | | | | | | | |
| 4 | 0.04 | 0.00 | 0.02 | 0.05 | | | | | | | | | | | | | | | | | | | | |
| 5 | 0.08 | 0.00 | 0.05 | 0.05 | 0.07 | | | | | | | | | | | | | | | | | | | |
| 6 | 0.02 | 0.04 | 0.02 | 0.01 | 0.03 | 0.06 | | | | | | | | | | | | | | | | | | |
| 7 | 0.00 | 0.01 | 0.01 | 0.03 | 0.04 | 0.03 | 0.09 | | | | | | | | | | | | | | | | | |
| 8 | 0.06 | -0.02 | 0.03 | 0.03 | 0.05 | 0.05 | 0.01 | 0.06 | | | | | | | | | | | | | | | | |
| 9 | 0.05 | 0.01 | 0.06 | 0.05 | 0.05 | 0.03 | 0.04 | 0.04 | 0.04 | | | | | | | | | | | | | | | |
| 10 | 0.04 | 0.04 | 0.04 | 0.05 | 0.03 | 0.03 | 0.02 | 0.03 | 0.06 | -0.02 | | | | | | | | | | | | | | |
| 11 | 0.01 | 0.03 | 0.04 | 0.03 | 0.03 | 0.01 | 0.05 | 0.01 | 0.02 | 0.03 | 0.05 | | | | | | | | | | | | | |
| 12 | -0.01 | 0.06 | 0.05 | 0.02 | 0.00 | 0.00 | -0.02 | 0.02 | 0.01 | 0.02 | 0.02 | 0.02 | | | | | | | | | | | | |
| 13 | 0.04 | 0.03 | 0.01 | 0.02 | 0.01 | -0.01 | 0.04 | 0.01 | 0.02 | 0.04 | 0.03 | 0.01 | 0.12 | | | | | | | | | | | |
| 14 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.05 | 0.04 | 0.03 | 0.03 | 0.02 | 0.05 | 0.01 | 0.04 | 0.03 | | | | | | | | | | |
| 15 | 0.06 | -0.01 | 0.03 | 0.03 | 0.04 | 0.00 | 0.00 | 0.04 | 0.05 | 0.03 | 0.04 | 0.00 | 0.03 | 0.01 | 0.03 | | | | | | | | | |
| 16 | 0.03 | 0.02 | 0.03 | 0.05 | 0.03 | 0.05 | 0.01 | 0.05 | 0.04 | 0.04 | 0.02 | 0.03 | 0.00 | 0.02 | 0.03 | 0.16 | | | | | | | | |
| 17 | 0.03 | -0.01 | 0.04 | 0.05 | 0.06 | 0.02 | 0.02 | 0.05 | 0.05 | 0.04 | 0.03 | 0.01 | 0.02 | 0.06 | 0.08 | 0.05 | 0.01 | | | | | | | |
| 18 | 0.02 | 0.03 | 0.04 | 0.03 | 0.02 | 0.04 | 0.03 | 0.06 | 0.03 | 0.03 | 0.05 | 0.03 | 0.03 | 0.03 | 0.02 | 0.04 | 0.04 | 0.05 | | | | | | |
| 19 | 0.05 | 0.03 | 0.03 | 0.03 | 0.03 | 0.01 | 0.01 | 0.04 | 0.04 | 0.03 | 0.06 | 0.01 | 0.02 | 0.04 | 0.02 | 0.06 | 0.00 | 0.05 | -0.01 | | | | | |
| 20 | 0.01 | 0.03 | 0.04 | 0.02 | 0.00 | 0.02 | 0.01 | -0.01 | 0.05 | 0.04 | 0.06 | 0.02 | -0.01 | 0.03 | 0.05 | 0.01 | 0.02 | 0.03 | 0.05 | 0.25 | | | | |
| 21 | 0.03 | 0.03 | 0.03 | 0.03 | 0.01 | 0.02 | 0.01 | 0.03 | 0.01 | 0.04 | 0.03 | 0.03 | 0.02 | 0.04 | 0.01 | 0.01 | 0.00 | 0.03 | 0.02 | 0.01 | -0.09 | | | |
| 22 | 0.03 | -0.01 | 0.01 | 0.02 | 0.05 | 0.04 | 0.00 | 0.06 | 0.03 | 0.04 | 0.03 | 0.04 | 0.02 | 0.03 | 0.04 | 0.03 | 0.04 | 0.03 | 0.03 | 0.00 | 0.00 | 0.06 | | |
| 23 | 0.05 | 0.03 | 0.04 | 0.02 | 0.04 | 0.03 | 0.03 | 0.04 | 0.04 | 0.06 | 0.03 | 0.01 | 0.03 | 0.02 | 0.04 | 0.04 | 0.05 | 0.05 | 0.04 | 0.01 | 0.05 | 0.03 | -0.06 | |
| 24 | 0.00 | -0.01 | 0.01 | 0.07 | 0.04 | 0.03 | 0.02 | 0.06 | 0.03 | 0.03 | 0.01 | 0.04 | 0.02 | 0.04 | 0.00 | 0.02 | 0.04 | 0.04 | 0.03 | 0.02 | 0.00 | 0.03 | 0.05 | 0.06 |

Figure 2. Average Between and Within Item Set Correlations

The three alpha reliability coefficients, item $^{s}x$, testlet $^{s}x$, and mixed $^{s}x$, were compared to the generalizability coefficient to determine the effect of LID and testlet-based scoring on the measure of reliability.

## Results

### Research Question 1

To answer the question "*To what extent do the context-dependent item sets used* *in the Physics 105 final examination possess LID?"* the average between-set and within-set correlation was calculated for each item set. These average correlations are displayed in Figure 2. To identify the item sets that contain LID a cut-off threshold was established using the between-set correlation frequency distribution. Theoretically, any within-set correlations greater then the between-set correlations would be identified as LID. To ensure that item sets that contain LID were not missed, a probability of a Type I Error was set. A Type I error is not identifying an item set as containing LID when it does contain LID. In research, the probability of a Type I error is typically set at 5%. For this study the Type I error was set by referring to the cumulative frequency distribution of between set correlations shown in Table 2.

Table 2.
Between-Set Correlation Distribution

| Correlation Intervals | Frequency | Percentage | Cumulative Frequency | Cumulative Percentage |
|---|---|---|---|---|
| -.030 to -.021 | 1 | 0.36% | 1 | 100.00% |
| -.020 to -.011 | 6 | 2.16% | 7 | 99.64% |
| -.010 to -.001 | 11 | 3.96% | 18 | 97.48% |
| .000 to .009 | 21 | 7.55% | 39 | 93.52% |
| .010 to .019 | 47 | 16.91% | 86 | 85.97% |
| .020 to .029 | 56 | 20.14% | 142 | 69.06% |
| .030 to .039 | 59 | 21.22% | 201 | 48.92% |
| .040 to .049 | 43 | 15.47% | 244 | 27.70% |
| .050 to .059 | 23 | 8.27% | 267 | 12.23% |
| .060 to .069 | 9 | 3.24% | 276 | 3.96% |
| .070 to .079 | 2 | 0.72% | 278 | 0.72% |

A total of 3.96% of the between-set correlations are above the threshold of .06. For this study we had a 3.96% probability of committing a Type I error; it was as close to 5% that could be achieved with this data set, using the specified correlation intervals. The cutoff value of .06 (with a Type I error of about 4%) was used to determine if there is statistical LID in the within-item set correlations.

*How many of the 24 item sets in the test show evidence of LID?* Ten of the 24 within set correlations were equal to or greater than .06 (see Table 3). These correlations indicate that the 10 item sets contain LID. Note that two of the item sets, numbers 2 and 20, are unusually large.

Table 3.
Within-Set Correlations

| Item Set | Correlation |
|---|---|
| 1 | 0.093 * |
| 2 | 0.217 * |
| 3 | -0.008 |
| 4 | 0.046 |
| 5 | 0.074 * |
| 6 | 0.057 |
| 7 | 0.092 * |
| 8 | 0.064 * |
| 9 | 0.042 |
| 10 | -0.017 |
| 11 | 0.047 |
| 12 | 0.023 |
| 13 | 0.124 * |
| 14 | 0.032 |
| 15 | 0.026 |
| 16 | 0.157 * |
| 17 | 0.007 |
| 18 | 0.054 |
| 19 | -0.008 |
| 20 | 0.249 * |
| 21 | -0.091 |
| 22 | 0.060 * |
| 23 | -0.060 |
| 24 | 0.062 * |

*Correlations Greater than .059

*What is the average amount of LID in the different item sets, and how does this amount vary from set to set?* The average amount of LID within the item sets was .056. The LID within the sets varied from negative correlations that ranged from -0.008 to -0.091 and positive correlations that ranged from .007 to .249 (see Figure 3).

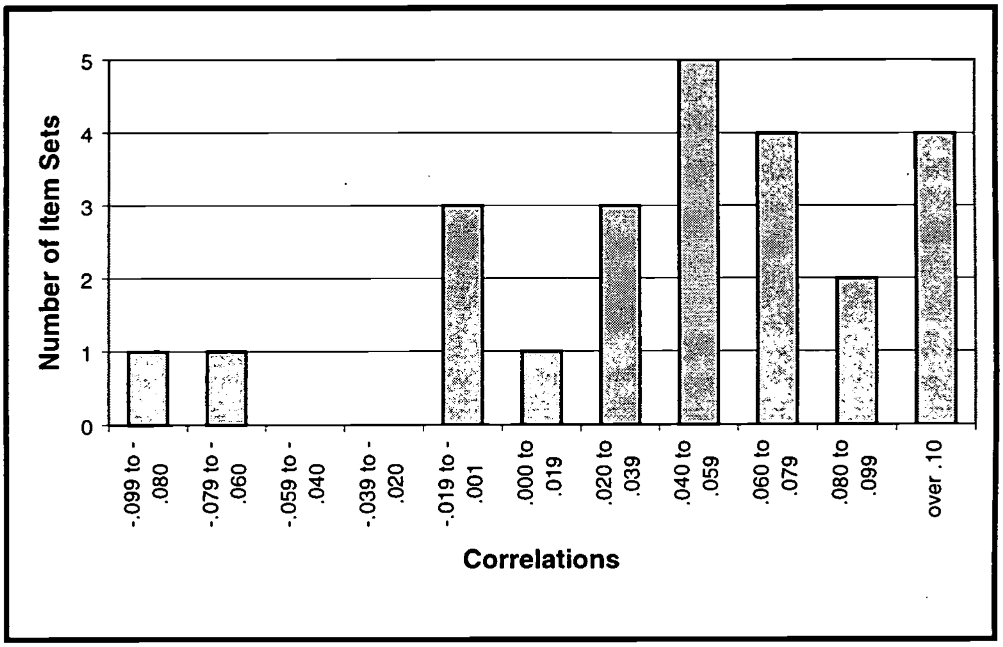


Figure 3. Local Item Dependence Within Item Sets

## Research Question 2

*To what extent does the presence of LID lead to an overestimate of the reliability of the scores obtained from the Physics 105 examination as evidenced by the difference in the magnitude of Cronbach's alpha for item ${}^{s}x$ and the generalizability coefficient?* To answer the second research question item ${}^{s}x$ was calculated as well as the G-coefficient for the test data. These reliability estimates where compared by subtracting the classical test theory reliability estimates from the G-coefficient. The results can be seen in Table 4. The item ${}^{s}x$ overestimated the reliability of the scores by .0204.

Table 4.
Reliability Estimates and Differences

| Type of | | Difference from | Standard Error |
|---|---|---|---|
| Estimate | Reliability | G-Coefficient | Measurement |
| Item $^{s}x$ | 0.9099 | 0.0204 | 1.2466 |
| Testlet $^{s}x$ | 0.8607 | -0.0287 | 1.5502 |
| Mixed $^{s}x$ | 0.8817 | -0.0077 | 1.4286 |
| G-Coefficient | 0.8895 | 0 | 1.3811 |

Research Question 3

*If testlet-based scoring is used for all of the item sets including both the sets that show evidence of LID and those that do not posses LID, how closely does the Cronbach's alpha for testlet* $^{s}x$ *match with generalizability coefficient?* The testlet $^{s}x$ underestimated the reliability of the scores by .0287. When compared to the G-coefficient, the testlet $^{s}x$ appeared to be a less accurate estimate then the item $^{s}x$.

*If testlet-based scoring is used only for item sets that show evidence of LID and item-based scoring is used for item sets that do not contain LID, how closely does the Cronbach's alpha for mixed* $^{s}x$ *match the generalizability coefficient?* The mixed $^{s}x$ only underestimated the reliability by .0077. A mixed $^{s}x$ results in an estimate of the reliability that is the closer to G-coefficient than item $^{s}x$ and testlet $^{s}x$.

## Discussion

Identifying LID

The results of this study provide evidence that it is possible to write context-dependent items that do not contain LID. The Physics 105 test consisted of 24 context-dependent item sets, with 100 items. Only 10 of those context-dependent item sets exhibited evidence of LID. With careful thought and time, context-dependent item sets can be written or rewritten to successfully test higher order thinking skills without LID. However, if a test is constructed to contain LID or it can not be re-administered with rewritten items then there are ways to score the test that will reduce the effects of LID.

## Assessing the Effects of LID and the Advantages of Using Testlet Scoring

The traditional item-based alpha coefficient overestimates the reliability of scores obtained from a test composed of context-dependent item sets that contain LID. Hence, a teacher or researcher attempting to make decisions based on scores from context-dependent items should be cautious about using classical reliability coefficients because their use may lead to erroneous decisions or interpretations.

On the other hand, the use of the testlet-based alpha coefficient underestimates the reliability of a test composed of context-dependent items sets that contain LID. Using testlet scoring produces a reliability estimate that is even less accurate than using individual item scores. Testlet-based reliability estimates for tests with context-dependent items that contain LID should also be used with caution because they underestimate the reliability and may also lead to inaccurate conclusions or decisions.

A traditional coefficient alpha computed for a mixture of testlets and individual item sets on a test composed of context-dependent item sets results in a reliability estimate that falls in between the individual item and testlet estimates as was expected. The resulting estimate more Closely approximates the G-coefficient, which is a more accurate estimate of the reliability of the scores. The mixed coefficient only underestimated the reliability by .0077. This is the most accurate coefficient alpha estimate that we computed.

Is the difference in the reliability estimates practically significant? Is the magnitude of the differences such that decisions will be affected by an under or overestimation of LID? To answer a question of practical significance we will need to look at the how test scores are interpreted. One of the problems associated with the use of test scores is that test users sometimes interpret small differences in examinee's observed scores as being indicative of real differences in the trait which the test is designed to assess. To prevent this kind of misinterpretation, responsible test users are advised to use the SEM as a means of estimating the amount of error contained in an individual examinee's observed score. The size of the SEM for any set of test scores is inversely related to the size of the reliability coefficient. Consequently, if an overestimate of the reliability coefficient is used then the SEM will be underestimated. Conversely, if an underestimate of the reliability coefficient is used then the SEM will be overestimated. Refer back to Table 4 for the reliabilities and SEM of the reliabilities.

The most accurate SEM is estimated from the G-coefficient. The item $^{S}x$ reliability estimate produces a SEM that is 9.74% greater than the SEM from the G-coefficient. The testlet reliability estimate produces a SEM that is 12.25% less than the SEM from the G-coefficient. The mixed $^{S}x$ reliability estimate produces a SEM that is only 3.44% less than the G-coefficient. It is the most accurate, as was expected.

One use of the SEM is to calculate confidence bands for individual student scores, see the

equation below:

$$C.I. = O_i \forall 2 \text{ x } SEM \,, \tag{3}$$

where $O_i$ represents the observed test score for an individual student. When confidence bands are computed the effect of an overestimated or underestimated SEM is magnified by 2. A SEM that is overestimated produces an extra wide confidence interval. In this study an overestimated SEM increases the confidence band by 18.8%. When confidence bands are too wide, then the intervals overlap and students are interpreted as being the same when they are really different. A student could be given a grade that is an overestimate of their true score. This would be a false positive result. A SEM that is underestimated produces a narrow confidence interval. In this study, an underestimated SEM produces a confidence band that is 24.6% narrower. When confidence intervals are too narrow, then the intervals are separated and students are interpreted as being different when they are really the same. A student could be given a score that underestimates their true score. This is a false negative result. Clearly, the reliability affects the SEM which is used to calculate confidence bands, and therefore influences decisions. In this study, using a mixed-scored reliability estimate results in a SEM that produces a confidence band that is only 6.5% wider. An accurate estimate of reliability is essential when using test scores to make valid, dependable decisions.

For example, in this study, if individual item or testlet scoring is used to score the final examination, in which context-dependent items that contain LID are used, the reliability will be either over or underestimated, depending on the scoring method. Consequently, the Physics 105 students could be given a lower or higher test score than they earned. An inaccurate final test score leads to an inaccurate course grade. A majority of the students enrolled in Physics 105 are in pre-professional programs, such as a program for medical school. Acceptance into medical school is an extremely competitive endeavor. A course grade could influence the decision to accept or deny a student. An inaccurate course grade could lead to a wrong decision as to which students will be accepted. A student that does not qualify, but was given an overestimated grade could be falsely accepted and a student that does qualify, but was given an underestimated grade may be falsely declined. Wrong decisions can be avoided by using reliable scores. In this case the scores given on a final examination can impact a life-changing decision. It is necessary to have reliable test scores when making test- or grade- based decisions.

The conclusion that is drawn from these results is that tests with context-dependent items should be used with caution. The scoring procedure and the method for estimating a reliability should be carefully considered. Scoring the item sets as individual items or as testlets is not the best solution. Both procedures produce inaccurate estimates of the reliability. However, the benefits gained from using the testlets, without the disadvantage can still be accrued by using mixed scoring accompanied by a mixed $^{s}x$ procedure for estimating reliability. The first step is to identify which item sets contain LID. Those sets that do contain LID should be scored as a testlet. Item sets that do not contain LID should not be scored as testlets, but should be scored as individual items. The reliability of the scores should be computed using a mixed $^{s}x$ coefficient.

The mixed coefficient produces a reliability estimate that is more accurate. It minimizes the lost item information, results in a longer test, and has lower error correlations than exclusive use of either testlet scoring or item scoring. A more reliable score results in an accurate SEM, which lead to good and accurate decisions. When using a mixed $^{s}x$ full advantage can be taken of the extremely versatile, efficient, and effective context-dependent item sets. When scored appropriately context-dependent items can be the answer to the push for tests that will assess higher order thinking and go beyond mere recall.

## Limitations of the Study

This study has a couple limitations that should be considered when interpreting the results. The number of items in each item set varied between 3 and 5 items. The between and within item set averages were not weighted by the number of items. This may be a source of error in the item set average correlations. The error was considered to be negligible, but may be a limitation.

Another limitation of the study is in the correlation matrix method that was used to identify LID the probability for a Type I error was set at about 4% due to the data and the way in which it was grouped into intervals. The probability for a Type I error is usually set at 5% or higher. The probability of a Type I error set at 4% is low and could lead to a mistake of not identifying LID when it was present within item sets. A higher probability would give a lower cut off threshold, thus identifying more item sets as containing LID.

Future Research

There are many possibilities for further research on this topic. Other studies could be done to account for the limitations. How does choosing the alpha for a Type I error effect the identification of LID? Also the differences between estimated alpha reliabilities as well as the differences between the alpha reliabilities and the g-coefficient could be tested for statistical significance.

The results of the study demonstrated the practical difference in reliabilities. However, the magnitude of the differences in the reliabilities were not as great as expected. The Physics 105 exam was well written and resulted in a relatively high reliability. Could the small difference in reliabilities be the result of a highly reliable test? Is a test with a high reliability more robust to the negative effects of LID?

Studies on the other methods of identifying LID should be conducted. The results of each method could be compared to see if the methods are comparable in which item sets they identify as containing LID. Methods involving the use item response theory lead to questions of measurement that are not an issue nor can they be answered using non IRT methods. Does LID affect test information, characteristic functions, trace lines, and item discriminations? LID underestimates the SEM. To what degree is it underestimated? Is it statistically significant? And does the scoring procedures explored in this study apply to SEM as well?

A topic that was not discussed in depth in this study is the cause of LID. Types of questions, response format, speededness, and test wiseness, fatigue, practice, external assistance and interference are all possible contributors to LID. There are various studies that have been done and studies yet to be done that address these issues.

# References

Andrews, K. M. (1989, March). *Methods of estimating the reliability of scores from multiple true-false tests.* Paper presented at the annual Meeting of the National Council on Measurement in Education, San Francisco.

Baron, J. B. & Wolf, D. B. (1996). *Performance-based student assessment: Challenges and possibilities*. Ninety-fifth yearbook of the National Society for the Study of Education, Part 1. Chicago: University of Chicago Press.

Brennan, R.L. (1992). *Elements of generalizability theory*. Iowa City, IA: American College Testing.

Brennan, R. L. (1999). *Manual for urGENOVA* (Iowa Testing Programs Occasional Papers No. 46). Iowa City, IA: University of Iowa.

Crocker, L. M. & Algina, J. (1986). *Introduction to classical & modern test theory.* Fort Worth, TX: Harcourt Brace Jovanovich.

Ebel, R.L. (1951). Writing the test item. In E.F. Lindquist (Ed.), *Educational measurement* (pp.185-249). Washington, DC: American Council on Education.

Ferrara, S., Huynh, H., & Baghi, H. (1997). Contextual characteristics of locally dependent open-ended item clusters in a large-scale performance assessment. *Applied Measurement in Education*, *10*, 123-144.

Ferrara, S., Huynh, H., & Michaels, H. (1999). Contextual explanations of local dependence in item clusters in a large scale hands-on science performance assessment. *Journal of Educational Measurement*, *36*, 119-140.

Guilford, J. P.(1936). *Psychometric methods* (1st ed.). New York: McGraw-Hill.

Haladyna, T.M. (1992). Context-dependent item sets. *Educational Measurement: Issues and Practice*, *10*, 21-25.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 147-200). New York: Macmillian.

Lee, G. & Frisbie, D. A. (1999). Estimating reliability under a generalizability theory model for test scores composed of testlets. *Applied Measurement in Education*, *12*, 237-255.

Nunally, J.C, (1967). *Psychometric theory.* New York: McGraw-Hill.

Shavelson, R. J. & Webb, N. M. (1991). *Generalizability theory a primer.* Newbury Park, California: Sage.

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

Wesman, A. G. (1971). Writing the test item. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 81-129). Washington, DC: American Council on Education.

Wainer, H. & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-201.

Wainer, H. & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 24*, 22-29.

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

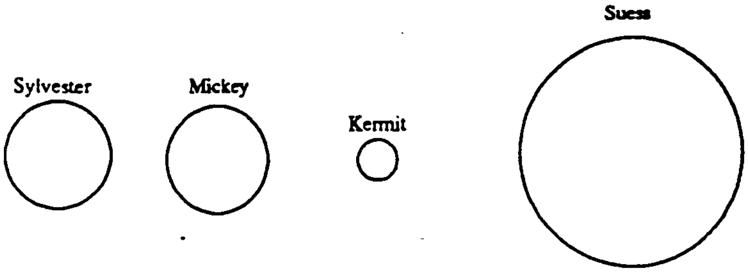
Worthen, B. R., White, K. R., Fan, X., & Sudweeks, R. R. (1999*). Measurement and Assessment in Schools* (2nd ed.). New York: Addison Wesley Longman, Inc.

Yen, W. M. (1993). Scaling performance assessments: strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-213.

Zimmerman, D.W. & Williams, R. H. (1980). Is classical test theory 'robust' under violation of the assumption of uncorrelated errors? *Canadian Journal of Psychology, 34*, 227-237.

The questions 71-73 refer to the information in the box below:

Suppose we have discovered a new solar system, comparable to our own, orbiting the star Kartune. It contains at least four planets (shown below-not necessarily in order or to scale). We have been able to obtain some of the information regarding this solar system. (m=mass in units of Earth masses ($M_e$), r=radius of planet in Earth radii ($r_e$), R=mean distance from center star in units of the Earth's orbit's radius ($R_e$))



Kartune:
m=$2.50 \times 10^{30}$ kg
r =$8.00 \times 10^{8}$ m

| Svlvester | Mickev | Kermit | Suess |
|---|---|---|---|
| m=2.0 $M_e$ | m=1.00 $M_e$ | m=1.00 $M_e$ | m=4.0 $M_e$ |
| r=1.0 $r_e$ | r=1.0 $r_e$ | r=0.5 $r_e$ | r=2.0 $r_e$ |
| R=1.0 $R_e$ | R=2.0 $R_e$ | R=4.0 $R_e$ | R=6.7 $R_e$ |

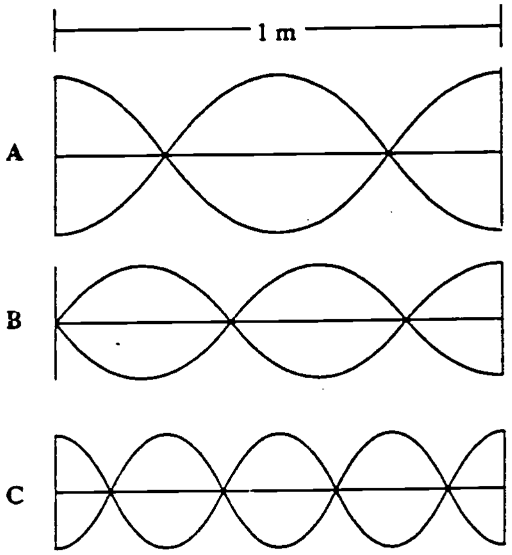
71. On the surface of which planet would you weigh the most?
   a. Sylvester
   b. Mickey
   .24 * c. Kermit
   d. Suess

72. Suppose Sylvester and Mickey are lined up on opposite sides of Kartune at equal distances. (Eg: ∘ O ∘ ) Consider only the forces acting *between the two planets*. Which planet has the greatest gravitational force acting on it?
   a. Sylvester, because it is more massive.
   b. Mickey, because it is less massive and therefore easier to be moved.
   c. Neither experiences a force because the forces add to zero.
   .41 * d. Neither, they both experience the same magnitude of force.

73. What would happen if Kartune suddenly disappeared, leaving its planets all by themselves in the solar system? Ignoring the gravitational forces between the planets, the planets would...
   a. continue in the same orbit because of Newton's first law.
   b. collapse toward the center where Kartune used to be.
   .92 * c. travel in a straight path outward, tangent to their orbits.
   d. stop orbiting and remain stationary relative to each other.

Questions 83-86 refer to the information and figure in the box below:

We have two organ pipes, both 1 m long, one closed at one end and one with both ends open. The figures below represent potential standing wave patterns for those pipes.



83. Which standing wave pattern is a possible pattern for the pipe closed at one end?
A. Wave A
.88 * B. Wave B
C. Wave C
D. Both A & C are possible.
E. All of A, B, & C are possible.

84. Which standing wave pattern is a possible pattern for the pipe open at both ends?
A. Wave A .83 * D. Both A & C are possible.
B. Wave B E. All of A, B & C are possible.
C. Wave C

85. What is the wavelength of the fundamental mode for the pipe open at both ends?
A. 0.25 m .47 * D. 2.00 m
B. 0.50 m E. 4.00 m
C. 1.00 m

86. What is the wavelength of the fundamental mode for the pipe closed at one end?
A. 0.25 m D. 2.00 m
B. 0.50 m .43 * E. 4.00 m
C. 1.00 m