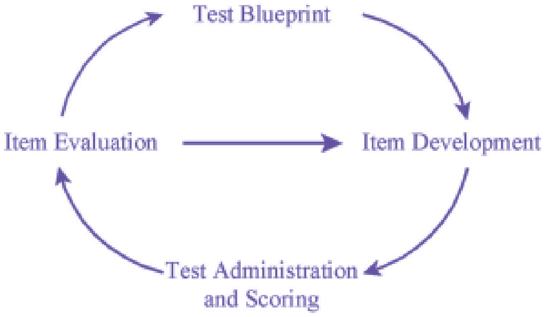
Helpful Tips for Creating Reliable and Valid Classroom Tests: Evaluating the Test

The first three articles in this series have provided a quick overview of the steps involved in developing effective classroom tests. The test development process begins with careful consideration of the exact content to measure and a decision on the types of items that will be used to measure it. Next is the item development stage, where actual test items are written. The test is then administered and scored. All too often, the scoring of tests and assigning of letter grades is viewed as the concluding step, the last piece of the puzzle. Although ultimately our job as instructors boils down to assigning grades to students, we do ourselves a disservice if we view assessment and test construction as a linear process. In fact, test development is a cyclical process; the data received from administering the test should be used to inform you about the appropriateness of the content and the effectiveness of the individual items in future exams. Although the students in your classes change semester to semester, assessment is ongoing. A model for thinking about the testing and assessment process is given below in Figure 1.

Figure 1. Testing process.



After we administer a test, we have a wealth of information about how students performed on each item. The most convenient way to organize all this information is in an item analysis (IA). An IA provides a breakdown of how different types of students performed on various aspects of each item. IAs are particularly useful for multiple-choice tests, but could conceivably be used for other item types as well. Instructors who bring their test data to Testing & Evaluation Services for scanning and scoring will receive a detailed IA report along with their scored rosters. The IA consists of two parts—a graph on the left side and a matrix of numbers on the right side—as shown in Figure 2.

(continued on page three)

The Learning Link is a quarterly newsletter published by the University of Wisconsin-Madison Teaching Academy. It aims to provide a forum for dialogue on effective teaching and learning. University teaching staff, including graduate teaching assistants and undergraduate students, are invited and encouraged to submit contributions.

Editors:
John DeLamater, Sociology
(608) 262-4357
delamate@ssc.wisc.edu

Mary Jae Paul
(608) 263-7748
mpaul@bascom.wisc.edu

Designer: Mary Jae Paul

UW-Madison Teaching Academy
133 Bascom Hall
500 Lincoln Drive
Madison, WI 53706
(608) 262-1677
http://wiscinfo.doit.wisc.edu/teaching-academy/

Chair:
Jay Martin, Mechanical Engineering
(608) 263-9460
martin@engr.wisc.edu

Project Assistant:
Elizabeth Apple
(608) 263-7748
EJApple@bascom.wisc.edu

UNIVERSITY OF
WISCONSIN
MADISON

Evaluating the Test

(continued from page two)

Figure 2. Sample item analysis for one item.

PERCENT RESPONDING CORRECTLY BY QUINTILE

```
5TH                                       *
4TH                              *
3RD                    *
2ND            *
1ST                               *
        |     |     |     |     |     |     |     |     |     |     |
        0     10    20    30    40    50    60    70    80    90    100
```

MATRIX RESPONDING BY QUINTILE

```
           A          B         C         D         E         O         M
5TH:       9          2         2         3         0         0         0
4TH:       7          1         6         3         0         0         0
3RD:       4          2         7         3         0         0         0
2ND:       2          6         7         2         0         0         0
1ST:       7          4         3         1         0         0         1
PROP:      [  0.35]   0.18      0.30      0.15      0.00      0.00      0.01
RPBI:      [  0.18]   -0.21     -0.07     0.11      0.00      0.00      -0.09
```

On the far left of both parts of the IA are the headings 5TH, 4TH, 3RD, 2ND, and 1ST. These correspond to quintile groupings of the examinees. To create these groupings, the total sample is rank-ordered based on their total score. Examinees with scores in the upper 20 percent are assigned to the 5TH quintile. The next-highest 20 percent are assigned to the 4TH quintile, and so on, with the bottom 20 percent of the examinees comprising the 1ST quintile. Hence, examinees are assigned to one of five groups, based on their total score.

The graph on the left hand side of Figure 2 plots the percentage of students in each of the five quintile groups answering the item correctly. Ideally, the five points in this graph will form a straight line with a positive, relatively flat slope (i.e., large jumps in percentage correct for each unit increase in quintile group). The picture is often not so clean, particularly when fewer than 100 examinees took the test. At a minimum, the points should look like they have a positive slope. If the set of points appears to have no slope (i.e., no relationship between quintile group and percentage correct scores) or a negative slope (i.e., higher quintile groups produced lower percentage correct scores), there is a good bet that the item is not functioning as intended.

The data matrix on the right hand side of Figure 2 is useful for refining the somewhat casual visual analysis that the graph provides. The data matrix is headed by a row of letters. A through E correspond to alternatives A through E for the item. Heading O stands for omit, and is used to tally the number of students who failed to provide an answer to the item. The heading for the final column, M, stands for multiple, and refers to situations where students recorded more than one answer for the item. The body of the matrix contains frequency counts of the number of examinees in each quintile group who selected each item alternative (or selected zero or multiple answers).

Quarterly Quote

"An education isn't how much you have committed to memory, or even how much you know. It's being able to differentiate between what you know and what you don't. "

- Anatole Feance

The last two rows of the data matrix correspond to two measures of item performance. The proportion of examinees selecting each response is presented in the column labeled PROP. The proportion selecting the correct answer, indicated in brackets, provides a measure of item difficulty. The item

(continued on page four)

Evaluating the Test

(continued from page three)

difficulty for the item given above is .35. Item difficulty varies from 0.00 to 1.00. For 4- or 5- choice multiple-choice items, an item is considered easy if it is answered correctly by more than 85% of the examinees, and is considered hard if it is answered correctly by less than 35% of the examinees. Because easy items tend to be answered correctly by nearly all students and hard items tend to be answered incorrectly for most students, except those who were able to correctly guess the answer, they are often unreliable. Reliability is improved by targeting the item difficulties between these two extremes.

The last row, labeled RPBI, indicates the correlation between whether a student selected a particular alternative (coded as a 1 if it was selected and a 0 otherwise) and the student's total score on the test. The statistic, called the point biserial correlation, associated with the correct answer provides a measure of item discrimination. In the above item, the discrimination is 0.18. Like all correlations, the RPBI varies from +1 to -1. For any given item, it is expected that students who did well overall should have answered the item correctly and students who did not do well should have answered the item incorrectly. To the extent that this pattern holds, the RPBI will be large and positive. If students doing well overall do poorly on an item, but students doing poorly overall tend to answer the item correctly, the RPBI will be negative. If there is no relationship between total score and item score, the RPBI will be near zero.

Large positive RPBIs for the keyed response are indicative of good items, because it means that the better achieving students are having success, while the lower achieving students are not. For classroom tests, RPBIs greater than or equal to 0.20 are usually fine, though higher values are better. RPBIs below 0.20 indicate a problem. Low positive RPBIs suggest that a student's overall level is hardly related to success on the item. Low positive RPBIs are often observed for very easy or very hard items. Negative RPBIs identify items on which low achieving students perform better than high achieving students.

Items with low positive or negative RPBIs should probably be either deleted from future exams or else revised. Careful study of the PROP and RPBI from the item distractors can be useful for identifying aspects of the item that may not be working well, and could potentially be changed to make the item function better. Whereas item keys are supposed to have moderate difficulties and high positive RPBIs, item distractors are supposed to have low-to-moderate difficulties and negative RPBIs. When item distractors have positive RPBIs, it suggests that something about the distractor is confusing, as good students are selecting it more often than weak students. If an item is in need of revision, a good place to start is by replacing any distractors that have positive RPBIs.

Another reason that items sometimes fail is that certain distractors are either selected by too many or not enough people. Occasionally, when common misunderstandings or misinterpretations are used as distractors, a distractor will be selected by an inordinately large percentage of students, perhaps more than selected the key. This, in and of itself, is not a problem, but to justify keeping a distractor that attracts over a third of the people (i.e., .33), the RPBI for that distractor must be negative and substantial (e.g., at least -.15). Finally, distractors selected by fewer than two percent of the students contribute very little to overall performance of the item, regardless of their RPBIs. Items that fail because they have one or two unattractive options can be improved by replacing those choices with more plausible alternatives.

To illustrate the item review process, the item in the sample IA is not working well. The item is pretty hard (p = .35), and its discrimination of 0.18 is below the threshold. Inspection of the distractor statistics shows that alternative B is working very well, as indicated by its strong negative RPBI. Alternative D, however, is quite confusing (its RPBI is +0.11), and should be replaced with something less attractive to good students. Alternative C has a negative RPBI, but its magnitude is very small, especially considering how many students selected C (PROP = .30). Therefore, it might be a good idea to replace C also.

It is very difficult to develop a test that yields good information about students' levels of understanding of the curriculum. This four-article series is aimed at providing some strategies for improving the quality of classroom tests. Clearly, these articles are not intended to provide an exhaustive description of test development guidelines, but are to be viewed as a starting point for improving classroom tests. For more information on test development, please check out Testing & Evaluation (T & E) Service's web site at http://www.wisc.edu/exams or call or come to T & E (373 Educational Sciences Bldg., 262-5863) and ask to talk with someone about help on developing classroom assessments.

James A. Wollack
Testing & Evaluation Services
UW-Madison, January 2004