Putting it all together: Producing Classroom Tests

Van Blerkom Chapter 10: Producing and Administering Written Tests

Incorporating aspects of the following points when developing tests can increase both efficiency and validity.

1. Number of Items to Include in a Test

(a) Speed versus Power Tests

With **speed** tests, the score is based upon the number of items that can be answered correctly within a given time limit. With **power** tests, one bases scores upon the number of items answered correctly without regard to time limits (within reason). Typically, less difficult, in fact very easy, items are used with speeded tests, but with power tests, difficult items are common. Speed and power can be viewed as two ends of a continuum, and classroom tests usually fall closer to the power end, but with recognition that time constraints must be present.

(b) Time Allotment

Time required to answer test items depends upon several factors such as content (e.g., a mathematical test may require more time to work than a spelling test) and skill level (e.g., higher-level cognitive activities often require more time to complete).

Some general rules for time allotment can be followed, depending upon the difficulty of each item:

* true-false – between 2 and 4 per minute;
* multiple-choice – between 1 and 2 per minute;
* short-answer – between 1 and 2 per minute; and
* brief-response essay – about 1 per 10 minutes (hopefully).

(c) Domain Emphasis

Emphasis, i.e., scoring weight per item, for each skill should be based on importance of each skill, not on length of time that is required to provide a response. Weighting of each skill should be assigned in one’s table of specifications.

(d) Obtaining Sufficient Accuracy

In short, the more test score accuracy desired (i.e., the better the evidence for validity and reliability), the more test items needed. It is very clear that as more items are included on a test, the accuracy of classification increases sharply (to a point, and then levels). Also, when one is testing over several content domains, it is important (for classification and remedial purposes) to include as many items on that domain as is possible. This is referred to as item sampling – ensuring enough items are included per domain to adequate assess that domain.

1. Item Difficulty

Item difficulty refers to the percentage of students who correctly respond to a given item. Thus, if an item is correctly answered by 57% of the students, it’s difficult is .57 or 57%.

(a) Item Difficulty and Achievement

For true-false and limited number of options multiple-choice items, item difficulty is usually best if between about 70 and 85%. Items that have a difficulty of 0 or 100% are not beneficial since they provide little to no information about student achievement, and therefore provide little information about remedial activity or alterations to classroom instruction. Thus, item difficulty is directly linked to the diagnostic role the items, and the test, are likely to assume.

(b) Item Difficulty and End-of-Term Grades

The relationship between item difficulty and course grades will be discussed in a later presentation.

1. Arranging Items Within a Test

(a) Ordering of Items

For power tests, the order of items does not matter although two options exist for ordering items: by content similarity for ease of discussing the test with the class, and by item similarity for ease of scoring.

(b) Physical Layout of Test Items

Items that are linked together should not be split over two pages. For example, items that require one to view a chart or options key should be on the same page as the chart or options key. Always proof-read tests multiple times.

(c) Test Instructions

Always include printed directions that are age appropriate for students. Instructions should be brief and concise, but thorough. Ideally instructions will include:

* how students should record their answers,
* how much time is allotted to the test,
* what to do when the test is completed,
* how answers are to be scored, and
* any special requirements, such as the use of scratch paper.

1. Developing a Classroom Test

Below are steps one should follow to develop tests that are content valid and produce reliable scores.

(a) Determine Purpose of Test Results

Will the test represent a formative or summative evaluation? If formative, then multiple items to measure well each skill or domain will be needed. If summative, fewer items per skill or content domain are typically used due to the broader scope of the test.

(b) Determine the Type and Number of Skills Assessed

Decide on the types of skills (e.g., application, analysis, etc.) one wishes to assess. Once this is determined, selecting item format will be easier since some are superior for measuring certain types of skills. Finally, the number of skills to assess also affects the item type to be used. For example, the greater then number of skills to assess, the less likely essay format can be used.

(c) Determine the Type and Number of Items to Be Used in this Test

Once step (b) is completed, one must next decide upon the specific number and type of items that will be used. This process will be made easier if it is integrated with step (d) below.

(d) Determine the Number of Items to be Associated with Each Objective or Devise a Table of Specifications

One should develop a table of specifications (or list of performance objectives) to facilitate the development of the test. Doing so will also help establish content validity for the test. Also, once all objectives are listed in the table, the task of determining the precise number of items and item types will be made easier to determine. Finally, one should also assign weight or point values for each objective and skill based upon the emphasis and importance of the given objective and skill.

(e) Prepare the Required Test Items

Now is the time to actually write the test items. Following a table of specifications will ease this process as it gives focus and direction for item writing. Evaluate each item against the skill and objective for which it was written and evaluate each item against its corresponding criteria found in presentations on each item type (i.e., multiple choice, true-false, short answer, essay). Should one use previously developed items, such as from earlier tests or textbook items, it will be important to evaluate these items against the criteria and table of specifications.

(f) Assemble the Items into a Test

When placing items into a test format, follow the guidelines provided in section 3 above “Arranging Items Within a Test.”

5. Establishing an Appropriate Testing Environment

The testing area should be familiar to students, quiet, and well lighted. Below are other considerations for testing environments.

(a) Interruptions

Eliminate, as far as possible, all distractions from the test. A common distraction is when an instructor offers addition testing instruction once the test is commenced. If students are to seek help or ask questions, it is best that this be done in an unobtrusive manner.

(b) Student Frustration

Without doubt, testing can be a frustrating experience. Some actions will exacerbate this situation, such as using tests as punishment, or having tests close to holidays. If possible, try to avoid these situations.

1. Correction for Guessing

While there are formulas for correcting for guessing, such corrections are not recommended for classroom tests.

Self-Test: Producing and Administering Written Tests

1. What percentage of students should be able to complete a classroom test within the time limits set by the teacher?

2. For each of the following formats, how many test items should students be able to answer in 30 minutes?

1. true-false
2. multiple-choice
3. completion, short-answer
4. brief-response essay

3. Is it appropriate to base the number of points assigned each test item on the time required to answer the item?

4. Does increasing or decreasing the number of test items improve the probability of correctly classifying a student?

5. Does a test item provide a more accurate indication of what the class in general has achieved or what each student has achieved?

6. In general, should a test item be included in the test if all students are expected to answer it correctly?

7. Do students score higher if easier items are placed near the beginning of the test?

For items 8 through 12, indicate (a or b) which activity should occur first when developing a written test.

8. (a) Determine how test results will be used, or (b) determine the type of skills the test will measure.

9. (a) Select the item format to be used, or (b) decide how many skills are to be tested.

10. (a) Determine the type of skills to be tested, or (b) select the item format to be used.

11. (a) Determine the number of items to be included in the test, or (b) determine the number of items to be used to measure each skill.

12. (a) Develop test items, or (b) determine how to use test results.

For items 13 through 19, use the following options to indicate which item format is being described.

1. Brief-essay
2. Completion or Short-Answer
3. Multiple-Choice
4. True-False

13. Using this format, a 20-item test would be expected to have the lowest reliability.

14. Using this format, a 30-minute test would be able to include the greatest number of items.

15. Problems in reliability of scoring are most significant for this item format.

16. Student would likely get the highest scores when items were written in this item format.

17. This item format is least able to measure a large number of objectives with a single test.

18. This item format is most able to measure a large number of objectives with a single test.

19. This item format is least able to ask students to evaluation alternate solutions to a problem.