Designing Multiple Choice Tests that Foster and Reflect Learning Outcomes

June 24-26, 2013
Assessment Retreat

Teresa Flateby, Ph.D.
tflateby@georgiasouthern.edu
Assessment Process

Select or develop measurable student learning outcomes

Make adjustments in curriculum strategies to address weaknesses and strengths

Academic Assessment Process

Identify classes, instructional strategies, assignments, and activities to foster outcomes

Select or develop measures that are able to provide information about outcome achievement and progress

Analyze assessment results

Measure student learning outcomes
Purposes of Multiple Choice Achievement Tests

• Measure individual student’s learning

• Evaluate students’ performance at program and institutional levels

• Use test results to identify and address gaps in learning

• Foster Course and Program Outcomes
Why Use Multiple-Choice Tests to Measure Achievement of Learning Outcomes?

• Efficient
  • More content coverage in less time
  • Faster to evaluate
  • Easier to evaluate test items
• In some cases, can provide a proxy to performance-based measures
Above All

Tests and the Assessment Process should **Promote Learning** and be able to **Accurately Measure** Desired Learning.
To Promote Learning and Provide Evidence of Outcome Achievement or Progress, Tests Must Be:

• **Valid**: Tests should be an accurate indicator of Content and Level of Learning/Outcome (Content Validity)

• **Reliable**: Tests should produce consistent results
Two Primary Test Development Theories

- Classical Measurement Theory
  - Test focus
  - Statistics produced are dependent upon sample taking test
- Item Response Theory
  - Item focus
  - Ability level of test takers and item difficulty are factors
Validity

• Tests must measure what you want your students to know and **be able to do with the content** (reach the cognitive demands specified in the outcomes).

• Tests must be consistent with instruction and assignments, which should foster the cognitive demands desired.
Process of Ensuring Validity

• Develop Table of Item Specifications also called Test Blue Print – necessary for developing effective achievement tests and guiding assessment

• Give test to a sample representative of the population

• Review item performance after administering test
Measuring Achievement of Outcomes with Multiple Choice Tests

- Create summative tests
- Develop **sets** of items to embed in courses indicating progress toward outcome achievement (formative)
- Develop course level tests that reflect program level outcomes

Validity applies to all
Test Blueprint Reflects Content and Cognitive Levels of Outcomes

<table>
<thead>
<tr>
<th>Content/ components of outcomes</th>
<th>Knowledge</th>
<th>Comprehension</th>
<th>Application and above</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Bloom’s Taxonomy of Educational Objective: Cognitive Domain

Useful for developing outcomes, test blueprint, and test items

- Analysis, Synthesis, Evaluation/Analyzing, Evaluating, Creating
  - Breaking down components of skill
  - Pulling it all together
  - Making judgment

- Application/Applying
  - Applying information/skill to new setting

- Comprehension (Lowest level of thinking)/”Understanding”
  - Paraphrasing and defining in own words

- Knowledge/Remembering
  - Recognizing
  - Recalling

Bloom, B.S., 1956, Madaus, et al., 1973
Institutional Outcome/Objective

• Students will demonstrate the critical thinking skills of analysis and evaluation in the general education curriculum and in the major.

Course Outcome

• At the completion of EDU 4015, students will analyze multiple choice tests by appropriately evaluating difficulty, discrimination indices, and distractors and will interpret the results for assessment purposes.
General Education Outcome

HISTORICAL AND CULTURAL PERSPECTIVES

Students are able to analyze artifacts, events, concepts or themes within the context of an evolving and diverse human experience

- World Literature – At the completion of the course, students will
  - Classify literary language, periods, and genres
  - Analyze literary texts in their cultural contexts
  - Differentiate the features of literatures produced by various world cultures
Constructing the Test Blue Print

1. List important course content or topics (at course level) or link to outcomes (program or course level).

2. Identify cognitive levels expected in outcomes.

3. Determine number of items for entire test and each cell based on: emphasis, time, importance, and testing time available, ability to help students generalize.
Base Test Blueprint on:

• Cognitive Level(s) of Outcome(s)
• Actual Instruction
• Classroom Activities
• Assignments
• Curriculum at the Program Level
• Assessment Purpose – Summative, Formative, or Both
# Test Blueprint of Questions and Levels for World Lit II

<table>
<thead>
<tr>
<th>Components</th>
<th>Knowledge Recall</th>
<th>Comprehension Skill/Concept</th>
<th>Application Strategic Thinking</th>
<th>Analysis Extended Thinking</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLO 1 Demonstrate familiarity with literary language, periods, and genres</td>
<td></td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>SLO 2 Analyze literary texts in their cultural contexts</td>
<td></td>
<td>2</td>
<td>0</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>SLO 3 Differentiate the features of literature produced by various world cultures</td>
<td></td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>TOTAL</td>
<td></td>
<td>9</td>
<td>6</td>
<td>10</td>
<td>25</td>
</tr>
</tbody>
</table>
Test Blueprint Activity

• Determine cognitive level of selected outcome

• Verify outcomes’ specificity

• Place into blueprint – indicate highest cognitive level

• Estimate number of items needed (or percentage of items)
Reliability: Repeatable or Consistent Results

- Not a characteristic of the test itself, but is based upon the specific sample of examinees.

- Typically based upon the correlation of two sets of scores - unrealistic in the classroom setting (e.g. test-retest, parallel forms).

- Internal consistency is the typical method used to calculate reliability in the classroom.
"The reliability of any set of measurements is logically defined as the proportion of their variance that is true variance... We think of the total variance of a set of measures as being made up of two kinds of variance: true variance and error variance... The true measure is assumed to be the genuine value of whatever is being measured." (Guilford, 1965, p. 488). So,

\[
\text{Reliability} = \frac{\text{true variance}}{\text{observed variance}}
\]

and, when measures and errors are uncorrelated,

\[
\text{Observed variance} = \text{true variance} + \text{error variance}
\]
Guidelines to Increase Reliability*
Goal is to reduce error variance

• Develop longer tests with well-constructed items.

• Make sure items are positive discriminators; students who perform well on tests generally answer individual questions correctly.

• Develop items of moderate difficulty; extremely easy or difficult questions do not add to reliability estimations.

• Administer test to a heterogeneous group of students.

*Guide for Writing and Improving Achievement Tests
Multiple Choice Items

- Follow Test Blueprint when developing items

- Ensure that items address cognitive level indicated in blueprint (map item numbers to Blueprint)

- Follow item-writing guidelines (see handouts)
### Breakdown by Question and DOK Level

<table>
<thead>
<tr>
<th>SLO 1</th>
<th>SLO 2</th>
<th>SLO 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Question</strong></td>
<td><strong>DOK Level</strong></td>
<td><strong>Question</strong></td>
</tr>
<tr>
<td>1</td>
<td>comprehension</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>comprehension</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>analysis</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>comprehension</td>
<td>10</td>
</tr>
<tr>
<td>9</td>
<td>application</td>
<td>11</td>
</tr>
<tr>
<td>12</td>
<td>application</td>
<td>22</td>
</tr>
<tr>
<td>13</td>
<td>comprehension</td>
<td>23</td>
</tr>
<tr>
<td>14</td>
<td>application</td>
<td>24</td>
</tr>
<tr>
<td>15</td>
<td>application</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>comprehension</td>
<td></td>
</tr>
</tbody>
</table>
Item Development Activity

• Develop two items for a row from your Test Blueprint at Comprehension or above
  • Use examples, Bloom’s Verbs, and item writing guidelines
  • Have your colleagues at your table provide feedback
After the Test Is Administered: Evaluate Test Results

1. KR-20: An outcome of .70 or higher. (Range 0-1.00)

2. Item discriminator: Point Biserial Correlation (-1.0+-1.0) should be positive; prefer .1 or higher, .3 and above is desired
   • Use when want a spread of scores, reflecting differences in student achievement
   • When question is “Are the right people getting the items right?”

S. Matlock-Hetzel, 1997
After the Test Is Administered: Evaluate Test Results cont’d

3. Difficulty Index; P-Value. (0-1.0) – Proportion answering item correctly

4. Analysis of Distractors – Percentage of students responding to incorrect responses
Use Results for Assessment Purposes

• Analyze performance on each item according to the outcome evaluated.

• Determine reasons for poor testing performance.
  • Faulty Item
  • Lack of Student Understanding

• Make adjustments to remedy these problems.
Item Analysis Activity

• Review Table of Item Analysis

• At your table, discuss the quality of the items

• Determine which items should be addressed and make suggestions for remedying any problems.
Guidelines for Developing Effective Items

Resources

• In *Guide for Improving Classroom Achievement Tests*, T.L. Flateby available online
• *Assessment of Student Achievement*, 2008, N.E. Gronlund Allyn and Bacon
• *Developing and Validating Multiple-Choice Test Items*, 2004, Thomas Haladyna; Lawrence Erlbaum Associates
• “KR-20 / Cronbach Alpha or Rasch Person Reliability: Which Tells the ‘Truth’?” [http://www.rasch.org/rmt/rmt113l.htm](http://www.rasch.org/rmt/rmt113l.htm)
• Additional articles and booklets are available at [http://fod.msu.edu/OIR/Assessment/multiple-choice.asp](http://fod.msu.edu/OIR/Assessment/multiple-choice.asp)
Internal Consistency for Dichotomous Measures
Kuder Richardson – 20 (KR-20)

\[ y = \frac{k}{(k - 1)} \left( 1 - \frac{\sum p q}{\sigma^2} \right) \]

- \( k \) = Number of Test Items
- \( p \) = Proportion of Correct Responses to a Test Item
- \( q \) = Proportion of Incorrect Responses to a Test Item
- \( \sigma^2 \) = Variance
Appendix

• Classical and Item Response Theories

• Outcomes Written for Each Taxonomic Level

• Items Written for Each Taxonomic Level
Classical and Item Response Theories

• Classical
  • Observed score and error score = True score

• Item Response
  • Item characteristic functions that are derived from a test are used to predict examinee scores at a specific ability level
Advantages

- Classical
  - Fewer assumptions have to be met
  - Smaller numbers of examinees are needed
  - Statistics used are more widely understood

- Item Response
  - Error can be specified
  - Difficulty of an item is independent of specific test takers
When appropriate

• Classical
  • If sample used to collect statistics on is similar to the population of students one making statements about
  • When the test will be administered to smaller numbers of students (fewer than 200-500)
  • Level of psychometric resources is limited

• Item Response
  • When ability level of students is a factor
  • When need to generalize to a wider population
  • When testing sample size is large
Outcomes Written for Each Taxonomic Level

1. Students will recall the four major food groups without error. (Knowledge)

2. Students will identify test construction steps to ensure test validity. (Knowledge)

3. By the end of the semester, students will summarize the main events of a story in grammatically correct English. (Comprehension)

4. Students will correctly identify if examples represent procedures that affect validity, reliability or both. (Comprehension)
5. Students will choose test construction scenarios that affect validity or reliability. (Application)

6. Given a presidential speech, students will be able to point out the positions that attack a political opponent personally rather than the person’s political programs. (Analysis)

7. Given an Item Analysis table, students will correctly interpret data contained in the table. (Analysis)
8. Students will describe the interrelationships among acts in a play. (Analysis)

9. Given a short story, students will write a different but plausible ending. (Synthesis)

10. Given a description of a country’s economic system, students will defend it by basing arguments on principles of socialism. (Evaluation)

Reference:

Examples for items written at the knowledge level

• *Outcome: Identifies the meaning of a term.*

• Reliability is the same as:
  • **A. consistency.**
  • B. relevancy.
  • C. representativeness.
  • D. usefulness.

• *Outcome: Identifies the order of events.*

• What is the first step in constructing an achievement test?
  • A. Decide on test length.
  • **B. Identify the intended learning outcomes.**
  • C. Prepare a table of specifications.
  • D. Select the item types to use.
Develop Comprehension Items

• Modify existing items (or items from your test) to reflect Comprehension level

• Develop Comprehension items using the examples below as a guide
Comprehension item suggestions cont.

• Comprehension can be evaluated by determining if students can **interpret if a statement or example** is consistent with a concept, principle, procedure or rule.

For example, the stem of an item could ask:

• **Which of the following statements is an example of a democratic political belief?** None of the distractors or the correct response should have been discussed in class or read in the course materials to ensure that students have at least a basic understanding of a democratic political belief.

http://jonathan.mueller.faculty@noctrl.edu/toolbox/tests/
Modify Knowledge questions to make them Comprehension or Application levels

Change:
The first stage of alcoholism is characterized by
a. malnutrition
b. addiction to alcohol
c. rationalization of drinking behavior*
d. reverse alcohol tolerance

To:
Susan believes her drinking behavior will lessen once she finishes a big project. Susan's explanation is particularly representative of the
a. first stage of alcoholism*
b. second stage of alcoholism
c. third stage of alcoholism
d. fourth stage of alcoholism

Answering the first form of the question correctly ("c") requires that students have memorized or can recognize the characteristics of the first stage. To require that students actually comprehend what those characteristics mean, the item can use an example such as the second question listed.

http://jonathon.mueller.faculty@noctrl.edu/toolbox/tests/
Comprehension Examples

• **Outcome: Identifies an example of a term.**
  • Which one of the following statements contains a **specific** determiner?
    • A. America is a continent.
    • B. America was discovered in 1492.
    • C. America **has some big industries.**
    • D. America’s population is increasing

• **Outcome: Interprets the meaning of an idea.**
  • The statement that “test reliability is a necessary but not sufficient condition of test validity” means that:
    • A. a reliable test will have a degree of validity.
    • B. a **valid test will have a degree of reliability.**
    • C. a reliable test may be invalid and a valid test unreliable.
Comprehension Examples cont.

- **Outcome: Identifies an example of a concept or principle.**
  - Which of the following is an example of a criterion-referenced interpretation?
    - A. Derik earned the highest score in science.
    - B. Erik completed his experiment faster than his classmates.
    - C. Edna’s test score was higher than 50 percent of the class.
    - **D. Tricia set up her laboratory equipment in five minutes**

- **Outcome: Predicts the most probable effect of an action.**
  - What is most likely to happen to the reliability of the scores for a multiple-choice test, where the number of alternatives for each item is changed from three to four?
    - A. It will decrease.
    - **B. It will increase.**
    - C. It will stay the same.
    - D. There is no basis for making a prediction
Develop Application Items

• Modify existing lower level items (either from above or your own items) to reflect Application level
• Determine if students can use what they have learned in a situation or example they have not encountered
• Use the examples below as a guide
Application Examples

Outcome: Distinguishes between properly and improperly stated outcomes.

Which one of the following learning outcomes is properly stated in terms of student performance?

A. Develops an appreciation of the importance of testing.
B. Explains the purpose of test specifications.
C. Learns how to write good test items.
D. Realizes the importance of validity.

Outcome: Improves defective test items.

Directions: read the following test item and then indicate the best change to make to improve the item.

Which one of the following types of learning outcomes is most difficult to evaluate objectively?

1. A concept.
2. An application.
3. An appreciation.
4. None of the above.

The best change to make in the previous item would be to:
A. change the stem to incomplete-statement form.
B. use letters instead of numbers for each alternative.
C. remove the indefinite articles “a” and “an” from the alternatives.
D. replace “none of the above” with “an interpretation”
A researcher wants to determine if a moderate exercise program could help lower blood pressure in people suffering from high blood pressure. However, the researcher is concerned that subjects' blood pressure might just naturally lessen over time and, consequently, she would not be able to tell if it was the result of the exercise program or not. To more accurately determine if the exercise program and not just time, is contributing to a reduction in blood pressure, the researcher should

a. establish a control group
b. extend the exercise program for a longer period of time
c. periodically check to see if the subjects are following the exercise program
d. compare the subjects to people without high blood pressure at the end of the study

http://jonathan.mueller.faculty@noctrl.edu/toolbox/tests/
Developing Analysis and Evaluation Items

• Examples, scenarios, diagrams and tables can be used to have students interpret, analyze or draw conclusions.
Directions: Read the following comments a teacher made about testing. Then answer the questions that follow by circling the letter of the best answer.

“Students go to school to learn, not to take tests. In addition, tests cannot be used to indicate a student’s absolute level of learning. All tests can do is rank students in order of achievement, and this relative ranking is influenced by guessing, bluffing, and the subjective opinions of the teacher doing the scoring. The teacher-learning process would benefit if we did away with tests and depended on student self-evaluation.”
Outcome: Recognizes unstated assumptions.

1. Which one of the following unstated assumptions is this teacher making?
A. Students go to school to learn.
B. Teachers use essay tests primarily.
C. **Tests make no contribution to learning.**
D. Tests do not indicate a student’s absolute level of learning.

Outcome: Identifies the meaning of a term.

2. Which one of the following types of test is this teacher primarily talking about?
A. Diagnostic test.
B. Formative test.
C. Pretest.
D. Summative test.
Analysis Example

Read carefully through the paragraph below, and decide which of the options A-D is correct.

“The basic premise of pragmatism is that questions posed by speculative metaphysical propositions can often be answered by determining what the practical consequences of the acceptance of a particular metaphysical proposition are in this life. Practical consequences are taken as the criterion for assessing the relevance of all statements or ideas about truth, norm and hope.”

A. The word “acceptance” should be replaced by “rejection.”
B. The word “often” should be replaced by “only.”
C. The word “speculative” should be replaced by hypothetical.”
D. The word “criterion” should be replaced by “measure.”

This question requires prior knowledge of and understanding about the concept of pragmatism.
The student is tested on his/her ability to analyze whether a word fits with the accepted definition of pragmatism.

*Gronlund, N.E., 1998, How to Make Achievement Tests and Assessments, Allyn and Bacon (pp. 42-46, 72-73.)
Analysis Example

Refer to the Item Analysis Table below when answering items 1-3 on the test:

<table>
<thead>
<tr>
<th>Item</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Prop. Passing</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>.08</td>
<td>0</td>
<td>.00</td>
<td>8</td>
<td>.10</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>.05</td>
<td>3</td>
<td>.04</td>
<td>67*</td>
<td>.84</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>18</td>
<td>.22</td>
<td>21*</td>
<td>.26</td>
<td>26</td>
<td>.32</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>.08</td>
<td>73*</td>
<td>.91</td>
<td>1</td>
<td>.01</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>.03</td>
<td>7</td>
<td>.09</td>
<td>2</td>
<td>.03</td>
<td>68*</td>
</tr>
<tr>
<td>6</td>
<td>43*</td>
<td>.54</td>
<td>1</td>
<td>.01</td>
<td>9</td>
<td>.11</td>
<td>26</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>.01</td>
<td>25</td>
<td>.31</td>
<td>1</td>
<td>.01</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>.05</td>
<td>9</td>
<td>.11</td>
<td>20*</td>
<td>.25</td>
<td>47</td>
</tr>
</tbody>
</table>

* Correct response
Use the table above to answer the following:

What is the most plausible interpretation of the results from item # 3 in the item analysis?

It is:

a) a difficult item.
b) an ambiguous item.
c) a well-performing item.
d) a norm-referenced item.
Evaluation Example

Judge the sentence in italics according to the criteria given below:

“The United States took part in the Gulf War against Iraq BECAUSE of the lack of civil liberties imposed on the Kurds by Saddam Hussein’s regime.”

A. The assertion and the reason are both correct, and the reason is valid.
B. The assertion and the reason are both correct, but the reason is invalid.
C. The assertion is correct but the reason is incorrect.
D. The assertion is incorrect but the reason is correct.
E. Both the assertion and the reason are incorrect.

A knowledge and understanding of Middle East politics is assumed. The student is tested in the ability to evaluate between cause and effect in the sentence in terms of predefined criteria.

http://www.uct.ac.za/projects/cbe/mcqman/mcqappc.html