Assessing Student Learning: Creating Good Test Items

Janette B. Benson, Office of Academic Assessment

Educational Objectives: What is *Bloom's taxonomy?*

Bloom's taxonomy is a classification system of **educational objectives** based on the level of student understanding necessary for achievement or mastery. Educational researcher Benjamin Bloom and colleagues have suggested six different cognitive stages in learning (Bloom, 1956; Bloom, Hastings & Madaus, 1971).

Bloom's cognitive domains are, in order, with definitions:

1. Knowledge	Involves the simple recall of information; memory of words, facts and concepts
2. Comprehension	The lowest level of real understanding; knowing what is being communicated
3. Application	The use of generalized knowledge to solve a problem the student has not seen before
4. Analysis	Breaking an idea or communication into parts such that the relationship among the parts is made clear
5. Synthesis	Putting pieces together so as to constitute a pattern or idea not clearly seen before
6. Evaluation	Use of a standard of appraisal; making judgments about the value of ideas, materials or methods within an area

There is an implied hierarchy to Bloom's categories, with **knowledge** representing the simplest level of cognition and the **evaluation** category representing the highest and most complex level. Instructors can identify the level of chosen classroom objectives and create assessments to match those levels. One can write items for any given level. With objectively scored item formats, it is fairly simple to tap lower levels of Bloom's taxonomy and more difficult, but not impossible, to measure at higher levels. By designing items to tap into teacher-chosen levels of cognitive complexity, classroom assessments increase validity.

Choosing the appropriate Bloom level for test items

Instructors choose the appropriate cognitive level for classroom objectives and a quality assessment is designed to measure how well those objectives have been met. Most items written by instructors and those in test banks packaged with textbooks are at the knowledge level. Most researchers consider this unfortunate because classroom objectives should be, and usually are, at higher cognitive levels than simply memorizing information. When new material is being introduced, however an assessment probably should include at least a check that basic new facts have been learned. The cognitive level of students, particularly their ability to think and understand abstractly and their ability to solve problems using multiple steps, should determine the best level for classroom objectives, and, therefore, the best level for test items. Researchers believe that teachers should test over what they teach in the same way that they teach it.

Knowledge **Comprehension**	**Application**	**Analysis** **Synthesis** **Evaluation**
Multiple Choice (MC)	MC	MC
True/False (TF)	Short Answer	Short Answer
Matching	Problems	Essay
Completion	Essay	
Short Answer	Performance	

How to write test items using Bloom's taxonomy.

Follow these guidelines to create items or tasks that require the type of thinking at each level of Bloom's taxonomy:

Cognitive Level	Test Item Example	Characteristics of Test Items
1. Knowledge	Who wrote The Great Gatsby? A. Faulkner B. Fitzgerald C. Hemingway D. Steinbeck	Requires only rote memory to answer correctly. Requires such skills as recall, recognition, repeating back.
2. Comprehension	What is a prehensile tail?	Includes phrases like *in your own words* and *what does this mean?* Requires such skills as paraphrasing, summarizing, and explaining.
3. Application	If a farmer owns 40 acres of land and buys 16 acres more, how many acres of land does she own?	Includes words like *use, do, modify, compute, produce.* Requires such skills as performing operations and solving problems.
4. Analysis	Draw a map of your house, identifying the location of each bedroom.	Includes phrases like *identify, break down, draw a diagram.* Requires such skills as outlining, listening, logic and observation.
5. Synthesis	Based on your understanding of the characters, describe what might happen in a sequel to *Flowers for Algernon.*	Includes words like *compare, describe, contrast, build.* Requires such skills as organization, design and creativity.
6. Evaluation	Which musical film performer was probably the best athlete? A. Maurice Chevalier B. Frank Sinatra C. Fred Astaire D. Gene Kelly	Includes phrases like *support, explain, apply standards, judge.* Requires such skills as making informed judgments, criticism, forming opinions.

References

Bloom, B.S. (Ed.). (1956). *Taxonomy of educational objectives: The classification of educational goals.* Handbook 1. *Cognitive domain.* New York: McKay.

Bloom, B.S., Hastings, J.T., & Madaus, G.F. (1971). *Handbook on formative and summative evaluation of student learning.* New York: McGraw-Hill.

Phye, G.D. (1997). *Handbook of classroom assessment: Learning, adjustment, and achievement.* San Diego, CA: Academic Press.

Designing Multiple-Choice Questions

A multiple-choice question is a type of item where students are presented with a question or instruction (a *stem*) and select the correct answer or response from a list of answer options. Technically, matching items, true-false items, and a variety of other specific item types where correct answers are available and students select the correct answer, are all multiple-choice questions.

1. Who wrote The Great Gatsby? **←Question Stem**

A. Faulkner ← **Distractor**
B. Fitzgerald ← ***Correct Answer ("Keyed Answer")***
C. Hemingway ← **Distractor**
D. Steinbeck ← **Distractor**

A few of the critical guidelines from those sources (Frey, Petersen, Edwards, Pedrotti, & Peyton, 2003; Haladyna & Downing, 1989a, 1989b; Haladyna, Downing & Rodriguez, 2002) are presented below.

Guideline 1.	**There should be 3 to 5 answer options.** Items should have enough answer options to make pure guessing difficult, but not so many that the distractors are not plausible or the item takes too long.
Guideline 2.	**"All of the Above" should not be an answer option.** Some students will guess this answer option frequently as part of a test-taking strategy. Other students will avoid it as part of a test-taking strategy. Either way, it does not operate fairly as a distractor. Additionally, to evaluate the possibility that "All of the Above" is correct requires analytical abilities which vary across students. Measuring this particular analytic ability is likely not the targeted goal of the test.
Guideline 3.	**"None of the Above" should not be an answer option.** This guideline exists for the same reasons as Guideline 2. Additionally, for some reason, teachers do tend to create items where "None of the Above" is the correct answer, and some students know this.
Guideline 4.	**All answer options should be plausible.** If an answer option is clearly not correct because it does not seem related to the other answer options, is from a content area not covered by the test, or because the teacher is obviously including it for humorous reasons, it does not operate as a distractor. Students are not considering the distractor, so a four-answer-option question is really a three-answer-option question and guessing becomes easier (i.e., 33% vs 25% chance correct just by guessing).
Guideline 5.	**Order of answer options should be logical or random.** Some develop a tendency to write items where a certain answer option (e.g. B or C) is correct. Students may either pick up on this, as part of a test-taking strategy, often guess B or C. You can control for any tendencies by placing the answer options in an order based on some rule (e.g. shortest to longest, alphabetical, chronological). Another solution is to scroll through the first draft of the test on their word processors and attempt to randomize the order of answer options.

Guideline 6.	**Negative wording should not be used.** Some students read more carefully or process words more accurately than others, and the word "not" can easily be missed. Even if the word is emphasized so no one can miss it, educational content tends not to be learned as a collection of non-facts or false statements, but, one would think, is likely stored as a collection of positively worded truths.
Guideline 7.	**Answer options should all be grammatically consistent with stem.** If the grammar used in the stem makes it clear that the right answer is a female or is plural, make sure that all answer options are female or plural.
Guideline 8.	**Answer options should not be longer than the stem.** An item goes more quickly if the bulk of the reading is in the stem, followed by brief answer options. A good multiple-choice question looks like this: 1. ========== A. ==== B. ==== C. ==== D. ====
Guideline 9.	**Stems should be complete sentences.** If a stem is a complete question, ending with a question mark, or a complete instruction, ending with a period, students can begin to identify the answer before examining answer options. Students must work harder if stems end with a blank or a colon or simply as an uncompleted sentence. More processing increases chances of errors.

Advantages of Multiple-choice Tests:

- test knowledge quickly within large groups
- be used to provide quick feedback
- be automatically scored
- be analyzed with regard to difficulty and discrimination, and
- be stored in banks of questions and re-used as required

Disadvantages of Multiple-choice Tests:

- take a lot of time to construct
- test knowledge and recall only
- never test literacy, or ability to analyze
- never test creativity, or unique thinking, and
- encourage students to take a surface approach to learning

What is a *matching item?*

Matching items are presented in groups as a series of stems or prompts that must be matched by the student to one of a group of possible answer options. The format is particularly useful when the objective to be measured involves association skills or the ability to recognize, categorize, and organize information. Matching items can be written to measure high levels of understanding but are most typically used at the knowledge level and for younger students.

Matching Items

Match each work with its author.
Answer options may be used more than once or not at all.

(Stems)	*(Answer Options)*
_____ 1. The Great Gatsby	A. Updike
_____ 2. The Grapes of Wrath	B. Salinger
_____ 3. The Sound and the Fury	C. Faulkner
_____ 4. Of Mice and Men	D. Fitzgerald
	E. Hemingway
	F. Steinbeck

Designing matching items

As with multiple-choice items, there has only been a small amount of empirical research on the characteristics of matching items and how they affect validity or reliability. In addition to research findings, there is also a common set of recommendations found in classroom assessment textbooks. A few of the critical guidelines from both these types of data (Frey, Petersen, Edwards, Pedrotti, & Peyton, 2003; Haladyna & Downing, 1989a, 1989b; Haladyna, Downing & Rodriguez, 2002) are presented below.

Guideline 1.	**There should be more answer options than stems.** As with many item-writing rules, the idea is to generate as many plausible answer options as possible, so students must have the knowledge to get the question correct.
Guideline 2.	**Answer options should be available more than once.** As with Guideline 1, this increases the number of functional distractors and increases the validity of the items. With this guideline, it is important that the instructions for the matching section indicate that answer options may be used more than once or not at all, so all students are aware of the rule.
Guideline 3.	**Directions should include basis for match.** A brief instruction identifying the category of stems and answer options (e.g. leaders and nations, species and phylum) helps students to focus on what constitutes a match, so they can concentrate on choosing the correct answer.
Guideline 4.	**Number of answer options should be < 17.** It is believed that college students can handle longer matching sections with many answer options, but too many options can slow down even the quickest of test-takers (especially when Guidelines 1 and 2 are followed). A well-made classroom assessment should not be exhausting for students.

Guideline 5.	**Matching stems should be on the left and answer options on the right.** Students are used to reading from left to right, and the process of matching two concepts together is similar to the construction and comprehension processes which occur when reading sentences.

What is item analysis?

Item analysis is a process of examining class-wide performance on individual test items. There are three common types of item analysis which provide teachers with three different types of information:

- ***Difficulty Index*** - Teachers produce a difficulty index for a test item by calculating the proportion of students in class who got an item correct. (The name of this index is counter-intuitive, as one actually gets a measure of how easy the item is, not the difficulty of the item.) The larger the proportion, the more students who have learned the content measured by the item.

- ***Discrimination Index*** - The discrimination index is a basic measure of the validity of an item. It is a measure of an item's ability to discriminate between those who scored high on the total test and those who scored low. Though there are several steps in its calculation, once computed, this index can be interpreted as an indication of the extent to which overall knowledge of the content area or mastery of the skills is related to the response on an item. Perhaps the most crucial validity standard for a test item is that whether a student got an item correct or not is due to their level of knowledge or ability and not due to something else such as chance or test bias.

- ***Analysis of Response Options*** - In addition to examining the performance of an entire test item, teachers are often interested in examining the performance of individual distractors (incorrect answer options) on multiple-choice items. By calculating the proportion of students who chose each answer option, teachers can identify which distractors are "working" and appear attractive to students who do not know the correct answer, and which distractors are simply taking up space and not being chosen by many students. To eliminate blind guessing which results in a correct answer purely by chance (which hurts the validity of a test item), teachers want as many plausible distractors as is feasible. Analyses of response options allow teachers to fine tune and improve items they may wish to use again with future classes.

Performing item analysis

Here are the procedures for the calculations involved in item analysis with data for an example item. For our example, imagine a classroom of 25 students who took a test which included the item below. The asterisk indicates that B is the correct answer.

	Number of Students Choosing Each Answer Option
Who wrote *The Great Gatsby*? A. Faulkner *B. Fitzgerald C. Hemingway D. Steinbeck	 4 16 5 0
Total Number of Students	25

Item Analysis Method	Procedures	Example
Difficulty Index- Proportion of students who got an item correct	Count the number of students who got the correct answer.	16 16/25 = .64

	Divide by the total number of students who took the test. Difficulty Indices range from .00 to 1.0.	
Discrimination Index- A comparison of how overall high scorers on the whole test did on one particular item compared to overall low scorers.	Sort your tests by total score and create two groupings of tests- the high scores, made up of the top half of tests, and the low scores, made up of the bottom half of tests. For each group, calculate a difficulty index for the item. Subtract the difficulty index for the low scores group from the difficulty index for the high scores group. Discrimination Indices range from -1.0 to 1.0.	Imagine this information for our example: 10 out of 13 students (or tests) in the high group and 6 out of 12 students in the low group got the item correct. High Group 10/13= .77 Low Group 6/12= .50 .77-.50=.27
Analysis of Response Options- A comparison of the proportion of students choosing each response option.	For each answer option divide the number of students who choose that answer option by the number of students taking the test.	Who wrote *The Great Gatsby*? A. Faulkner 4/25 = .16 *B. Fitzgerald 16/25 = .64 C. Hemingway 5/25 = .20 D. Steinbeck 0/25 = .00

Interpreting the results of item analysis

In our example, the item had a difficulty index of .64. This means that sixty-four percent of students knew the answer. If a teacher believes that .64 is too low, he or she can change the way they teach to better meet the objective represented by the item. Another interpretation might be that the item was too difficult or confusing or invalid, in which case the teacher can replace or modify the item, perhaps using information from the item's discrimination index or analysis of response options.

The discrimination index for the item was .27. The formula for the discrimination index is such that if more students in the high scoring group chose the correct answer than did students in the low scoring group, the number will be positive. At a minimum, then, one would hope for a positive value, as that would indicate that knowledge resulted in the correct answer. The greater the positive value (the closer it is to 1.0), the stronger the relationship is between overall test performance and performance on that item. If the discrimination index is negative, that means that for some reason students who scored

low on the test were more likely to get the answer correct. This is a strange situation which suggests poor validity for an item.

The analysis of response options shows that those who missed the item were about equally likely to choose answer A and answer C. No students chose answer D. Answer option D does not act as a distractor. Students are not choosing between four answer options on this item, they are really choosing between only three options, as they are not even considering answer D. This makes guessing correctly more likely, which hurts the validity of an item.

Research Articles

Frey, B.B., Petersen, S.E., Edwards, L.M., Pedrotti, J.T. & Peyton, V. (2003, April). *Toward a consensus list of item-writing rules.* Presented at the Annual Meeting of the American Educational Research Association, Chicago.

Haladyna, T. M. & Downing, S.M. (1989a). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2*(1), 37-50.

Haladyna, T. M. & Downing, S.M. (1989b). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2*(1), 51-78.

Haladyna, T.M., Downing, S.M., & Rodriguez, M.C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-334.

This handout was created by modifying information available online at
http://www.specialconnections.ku.edu/cgi-bin/cgiwrap/specconn/index.php