Creating quality multiple choice exams: Planning, developing, & improving

Cheryl Poth, PhD
poth@ualberta.ca
Bcissam El Hassar, MPP
Centre for Research in Applied Measurement & Evaluation

Among the issues for exam designing:

- Assessment method-level
  - Length and time allocated
  - Types of question formats
- Item-level
  - Unfamiliar vocabulary
  - Specific cultural references
  - Linking to previous experiences
  - Reading over-emphasized
- Others?

What guides our exam planning?

A. Big ideas
B. Item type
C. Exam emphasis

Overview of Exam design process
What guides our exam planning?

A. What are we assessing?
   ◦ What are the big ideas related to knowledge and skills that we have taught?

   ◦ This represents the learner outcomes that students would be expected to know/be able to do after the course is completed.

B. What can be assessed by a multiple-choice (MC) exam?
   ◦ What are some examples from the big idea activity that could be assessed using multiple choice?
   ◦ What might be other item types that can be used?
     - Ranking, matching, completion, short answer, long answer, performance assessment
   ◦ What are some examples from the big idea activity that would be better assessed using one of the other types?
   ◦ What can guide our choice of item focus?

Bloom’s Taxonomy

- Dr. Benjamin Bloom
- Creator of a Taxonomy of Educational Objectives

- Cognitive (mental skills)*
- Affective (attitudes and emotions)
- Psychomotor (physical skills)

Bloom’s Revised Taxonomy Verbs: Some Ideas

<table>
<thead>
<tr>
<th>Remembering: Remembering previously learned material</th>
</tr>
</thead>
<tbody>
<tr>
<td>check, read, identify, quote, name, recite, define, recognize, describe, list</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Understanding: Explaining ideas or concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>explain, illustrate, interpret, define, explain, match, model, memorize, translate, differentiate</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Applying: Using information in another familiar situation</th>
</tr>
</thead>
<tbody>
<tr>
<td>apply, complete, evaluate, use, create, design, formulate, predict, determine, classify, arrange</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Analyzing: Breaking information into parts to recognize understandings and relationships</th>
</tr>
</thead>
<tbody>
<tr>
<td>analyze, compare, contrast, classify, evaluate, hypothesize, critique, interpret, synthesize, point out</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Evaluating: Justifying a decision or course of action</th>
</tr>
</thead>
<tbody>
<tr>
<td>examine, locate, select, examine, criticize, judge, summarizing, evaluate, contrast, propose, recommend, validate, evaluate</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Creating: Generating new ideas, products, or ways of viewing things</th>
</tr>
</thead>
<tbody>
<tr>
<td>design, formulate, generate, organize, implement, plan, produce, develop, implement, modify, assume, evaluate, redesign, reorganize, operate, display</td>
</tr>
</tbody>
</table>

*Note: These verbs are not exhaustive and are provided as examples of the levels of the revised taxonomy.
How well does each Assessment Method Measure a Cognitive Level?

<table>
<thead>
<tr>
<th>Assessment Method</th>
<th>Multiple choice</th>
<th>Short/Long Answer</th>
<th>Performance Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Remembering</td>
<td>★★★</td>
<td>★★★</td>
<td>★</td>
</tr>
<tr>
<td>Understanding</td>
<td>★★★★★</td>
<td>★★★★★</td>
<td>★★★★★★</td>
</tr>
<tr>
<td>Applying</td>
<td>★★★★★</td>
<td>★★★★★</td>
<td>★★★★★★</td>
</tr>
<tr>
<td>Analyzing</td>
<td>★★★★★</td>
<td>★★★★★</td>
<td>★★★★★★</td>
</tr>
<tr>
<td>Evaluating</td>
<td>★</td>
<td>★</td>
<td>★★★★★★</td>
</tr>
<tr>
<td>Creating</td>
<td>★</td>
<td>★</td>
<td>★★★★★★</td>
</tr>
</tbody>
</table>

What guides our exam planning?

c. What does the exam emphasize?
   - Do you have a table of test specifications?
     - What is it?
     - Visual representation of the items in terms of both the content to be learned and the level of cognition expected of the students.
     - When to construct?
     - Ideally, prior to the beginning of instruction, but for sure before you finalize the exam.

EDPY 303 Midterm Exam Blueprint – Fall 2014

<table>
<thead>
<tr>
<th>Content area tested</th>
<th>Topics</th>
<th>Taxonomy levels tested</th>
</tr>
</thead>
<tbody>
<tr>
<td>Curriculum, Instruction, and Assessment Objectives</td>
<td>1-10</td>
<td>1-2</td>
</tr>
<tr>
<td>Bloom’s Taxonomy</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>Fair Assessment</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>Reliability and Validity</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Selected Response Items</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td>Completion Items</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Gathering Evidence of Learning</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>32</td>
<td></td>
</tr>
</tbody>
</table>

Course Content Level of Thinking – Blooms’ Taxonomy

<table>
<thead>
<tr>
<th>Topic</th>
<th>Remembering</th>
<th>Understanding</th>
<th>Applying</th>
<th>Analyzing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1-5, 6,7,10,13,14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>6,9,11,12,15-28, 31,32</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>39,40, 33-38</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>39,41-46</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>40,41-50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>20%</td>
<td>40%</td>
<td>30%</td>
<td>10%</td>
</tr>
</tbody>
</table>
Take home message for planning

- What and how you teach needs to be reflected in the assessments and linked with the intended big ideas.
- Bloom’s taxonomy is helpful for thinking about what you want students to be able to do and know.
- Exam blueprint/Table of test specifications is a useful tool to make explicit the emphasis of content and level of cognition.
- Questions:
  - How do you plan for your exams? Would Bloom’s taxonomy help improve your exam planning?
  - Would exam blueprint help improve the exam planning and help your students have clear expectations?

What guides our creation/selection of quality items?

A. Does the exam assess the depth of skills intended to assess?
  - Is there variety in the item level of cognition?

What level of Bloom’s do each of these instructional objectives require?

Students will:
1. Recall the basic purposes for commercial advertising
2. Describe the basic techniques advertisers use to sell products to consumers
3. Observe a series of television commercials and identify in each one the selling technique(s) employed
4. Differentiate the observed television commercials on the basis of their effectiveness in promoting a product/service
5. Script and perform a commercial designed to sell a product of their choice
6. Judge the effectiveness of the commercials created by their peers using a class-generated set of criteria

- A. Creating
- B. Evaluating
- C. Analyzing
- D. Applying
- E. Understanding
- F. Remembering
What guides our development/selection of technically sound items?

B. Do the items meet the criteria for being fair?
   - Are the items considered equitable?
     - Why is this important?
     - What might items that are not look like?
   - Are the items considered technically sound?
     - Why is this important?
     - What might items that are not look like?

An item is equitable when it is:
- free from racial, ethnic and sexual bias
- free of irrelevant material
- stated in appropriate and clear language
- free from pop cultural references that would not be familiar to all students

An item is technically sound when it is:
- the item is free of verbal clues to the answer
- the stem is focused on a single, meaningful problem
- key words in the stem are emphasized as needed
- distracters are all plausible
- alternatives are homogenous and parallel in structure
- alternatives are in some order that is logical and easily understood
- legitimate use of ‘all of the above’ and ‘none of the above’
- the item has correct spelling and grammar

What are the Parts of a MC Item?

Calculus was independently developed by
A. Barrow
B. Kepler
C. Leibniz
D. Pascal
Guidelines for a High-Quality MC Stem

- Focused wording: the item can be answered to some extent without looking at the alternatives
- Question or statement form
- Key words highlighted

Try to Avoid:
- Double-barreled stem (more than one idea)
- Verbal and grammar cues to the answer
- Content bias (e.g., references to pop-culture, gender bias)

Guidelines for a High-Quality MC Distractors

- Use common student errors
- Use language appropriate to the students
- Alternatives should all be homogeneous (length and complexity)
- Alternatives should be arranged in a logical order (e.g., pyramid or reverse-pyramid, alphabetical order, numerical order, etc.)

Try to Avoid:
- The 3:1 Split... One of the alternatives stands out “like a sore thumb”
- A multiple-choice item that is actually a “true-false” item because nearly all students can eliminate two distracters
- Unjustified use of “all/none of the above”

If you want to increase the cognitive level of your MC item—consider using a Source-Based M-C Item

Which of the following students deserve an A as their final grade?

A. Bob
B. Bob, Gwen and Roger
C. Bob, Gwen, Roger and Pam

This is the novel, yet familiar introductory (source) material.

Example: Improving Multiple-Choice Items

A table of test specifications
A. provides a more balanced sampling of content
B. specifies the method of scoring to be used on a test
C. indicates how a test will be used to improve learning
D. arranges the instructional objectives in order of their importance

This is one of a series of selected-response items that relates to the introductory material.
Example: How about this one?

The paucity of plausible, but incorrect, statements that can be related to a central idea poses a problem when constructing which one of the following types of test items?

A. essay items
B. true-false items
C. completion items
D. multiple-choice items

What guides our developing/selection of technically sound items?

C. What resources are available as a starting point?
- Do we have test bank items or old exams to adapt?
- What might be some of the adaptations you make?
  - Increase the level of cognition
  - Change item type
  - Reword to focus it on the “big idea”
- How will the exam be administered?

Learning Assessment Centre (LAC)

- Provides the possibility to administer secure digital/computerised exams
- Question format types supported are: MC, fill-in-the blank, short answer, essay.
- Benefits of digital exams are:
  - Multiple Choice and short answer questions can be instantly marked without using scantron sheets.
  - Students can use the accessibility features on the computer to accommodate vision difficulties and preferences.
  - SDSs students could write the exam with the rest of the students.
  - Instructors can access all of the exams in one place from flexible locations.
- Location: in 3-106 Education North
- Website: http://digital.ualberta.ca/learning-assessment-centre-lac

Take home message for developing MC questions

- Attend to the cognitive levels of items
- Make sure your items meet the criteria for being technically sound and equitable
- Consider differing ways of administering!
- Questions
  - Do you consider the cognitive levels of exam questions in your design?
  - Do you use other criteria for creating MC questions? How could the for mentioned check list help you with your question design?
Creating quality multiple choice exams

What guides our use of item analysis?

A. Exam reliability
B. Item difficulty
C. Item discrimination

Exam Reliability: Spot the Problem in the Following Alternative Response Item

Lengthening a test will increase its reliability.

What guides our use of item analysis?

A. Is the exam reliable?
   - If machine-scored, check KR-20
   - If hand-scored, look for patterns
   - What does this tell you?
   - What is considered high reliability vs. low reliability?

B. How difficult is the item?
   - If machine-scored, check DIF
   - What does this tell you?
   - Range from 0 to 1
     - Easy more than 0.9
     - Moderate 0.5 to 0.9
     - Difficult less than 0.5

ρ = # of correct answers/# of people taking the test
14. Elections are often held in non-democratic countries primarily as a means of
A. reinforcing the perceived legitimacy of the régime in power
B. providing an opportunity for citizens to effect political change
C. meeting the legal requirements imposed by legislated constitutions
D. providing the elite with an insight into popular attitudes and beliefs

ITEM 14: DIF = .628, RPB = .448, CRPB = .391 (95% CON = .286, 1.486)
RBIS = .572, CRBIS = .499, IRI = .216

<table>
<thead>
<tr>
<th>GROUP</th>
<th>N</th>
<th>NR</th>
<th>NF</th>
<th>O</th>
<th>1*</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOTAL</td>
<td>277</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>.628</td>
<td>.108</td>
<td>.130</td>
</tr>
<tr>
<td>HIGH</td>
<td>74</td>
<td>0</td>
<td></td>
<td></td>
<td>.905</td>
<td>.014</td>
<td>.000</td>
</tr>
<tr>
<td>MID</td>
<td>124</td>
<td>1</td>
<td></td>
<td></td>
<td>.629</td>
<td>.105</td>
<td>.161</td>
</tr>
<tr>
<td>LOW</td>
<td>79</td>
<td>0</td>
<td></td>
<td></td>
<td>.367</td>
<td>.203</td>
<td>.203</td>
</tr>
</tbody>
</table>

TEST SCORE MEANS
28.952  27.264  29.735  35.138
DISCRIMINATING POWER
- .130
- .330
- .078
.568
STANDARD ERROR OF D.P.
.011
.018
.013
.019

Question 14
This item meets all major standards. This question also discriminates very well. Note that although 20.3% of the lowest achieving group selected alternative C (3), none of the highest achieving group chose that same alternative.

What can you read from this analysis?
Why is any of this important?

- Validity of Interpretation
  - Accuracy
    - Extent to which an assessment method measures what we intend it to measure
  - Reliability of Assessment
    - Consistency of scores

Take home message for improving

- Item Analysis is an untapped source for information
- Important to consider measurement aspects of your items to ensure reliability in the exam and validity of your interpretations
- Questions:
  - Do you have previous exam' item analysis you can use to improve current tests?
  - How reliable are your exams? What steps could you take to improve the reliability of the result and validity of the results' interpretation?

References & Time for Questions

- Bloom’s taxonomy:
  http://www.odu.edu/educ/roverbau/Bloom/blooms_taxonomy.htm
- Fairness:
- Item Construction:
- Designing exams:
  http://www.ncsu.edu/felder-public/Papers/TestingTips.htm>