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

Cheryl Poth
cpoth@ualberta.ca
Katya Chudnovsky
Centre for Research in Applied Measurement & Evaluation
To what extent might these scenarios be familiar to you? (as a student or instructor)

What issues do these scenarios represent?
Among the issues are:

- Unfamiliar vocabulary
- Specific cultural references
- Linking to previous experiences
- Reading over-emphasized
- Others?
What our students are saying guides this workshop

- “Your exams are cover what we talked about in class.”
  - Focus on assessing what you taught
    - What assessment considerations must be embedded into the planning of an exam?
- “Your exams make me think harder than I have before.”
  - Focus on increasing quality of items
    - What principles guide the creation/selection of quality multiple choice items?
- “I feel your exams are fair.”
  - Focus on enhancing use of item analysis
    - How can the exam be improved for the future?

28 August 2012    Creating quality multiple choice exams
What guides our exam planning?

PLANNING

Big ideas

Content Balance

Item type
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
Read the list of twenty-one facts about the causes, course and consequences of the Second World War. Assume that within twenty years your students will remember only 5 facts. Place a check-mark beside the 5 facts you hope your students always remember.
Is there a commonality among the five facts you chose?
What guides our exam planning?

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, essay, performance assessment
     • What are some examples from the big idea activity that would be better assessed using one of the other types?
What guides our exam planning?

A. Does the exam balance the content intended to assess?

  ◦ 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
### Blueprint for an EDPY 303 Examination

<table>
<thead>
<tr>
<th>Topics</th>
<th>Selected Response</th>
<th>Constructed Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selected response=52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constructed response=3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Formative and Summative Assessments</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Curriculum, instruction, and assessment alignment</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Taxonomy Levels</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>Fair Assessment</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Reliability and Validity</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Selected Response</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>Developing Pencil &amp; Paper Tests</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Methods of Scoring &amp; Interpreting</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Assessment Audience</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Content areas tested.

Taxonomy levels tested.

Number of items at each taxonomy level for each content area.
What guides our creation/selection of quality items?

- Variety of cognitive levels
- Creation/Selection
- High quality items
- Resources

28 August 2012  Creating quality multiple choice exams
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?

Introducing Bloom’s Taxonomy!
What level of Blooms 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
Test-wiseness debrief

1. The purpose of the cluss in furmpaling is to remove
   A. cluss-prags
   B. tremalis
   C. cloughs
   D. plumots

2. Trassig is true when
   A. lusp trasses the vom
   B. the viskal flans, if the viskal is donwil or zortil
   C. the belgo frulls
   D. dissles lisk easily

3. The sigla frequently overfesks the trelsum because
   A. all siglas are mellious
   B. siglas are always votial
   C. the trelsum is usually tarious
   D. no trelsa are feskable

28 August 2012  Creating quality multiple choice exams 15
Test your test-wiseness!

4. The fribbled breg will minter best with an
A. Derst
B. Morst
C. Sorter
D. Ignu

5. Among the reasons for tristal doss are
A. The sabs foped and the foths tinzed
B. The kredges roted with orots
C. Few rakobs were accepted in sluth
D. Most of the polats were thonced

6. The mintering function of the ignu is most effectively carried out with
A. Raxma tol
B. The groshing stantol
C. The fribbled breg
D. A frally sush
What guides our creation/selection of quality items?

B. Do the items meet the criteria for being fair?
  ◦ Are the items considered high quality?
  ◦ Why is high quality important?
What are the Parts of a MC Item?

Calculus was independently developed by Newton and

A. Barrow
B. Kepler
C. Leibniz
D. Pascal

This is the stem.

These are the “alternatives”, “choices”, or “options”.

This is the “keyed response” or correct choice.

These are the “distracters”, (alternatives that are incorrect).
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)

➢ 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.

This is one of a series of selected-response items that relates to the introductory material.
Guidelines for a High-Quality Source-based MC Introductory Material

- relates to the content covered, **but** is new to the students
- Is brief and easy to read (appropriate level).
- Stimulates thought about an issue or topic
- Comes from a credible source and can be used (copyright!)
- Instructors can create their own introductory materials
- Follows the guidelines for MC item stems

- **Try to Avoid:**
  - Irrelevant source material
  - Testing knowledge directly cited in the source material
  - Testing trivia or memorized facts
  - Extraneous cues

28 August 2012  Creating quality multiple choice exams  22
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

Rewrite:

The main advantage of using a table of test specifications when preparing an achievement test is that it
A. improves the sampling of content
B. increases the objectivity of the test
C. reduces the amount of time required
D. makes the construction of test items easier

Guideline:

present a single clearly formulated problem in the stem and emphasize the key words
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

Guideline: write the stem of the item in simple, clear language

Rewrite:

A lack of plausible yet incorrect alternatives causes the greatest difficulty when constructing

A. essay items
B. true-false items
C. completion items
D. multiple-choice items
What guides our creation/selection of quality 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”
What guides our use of item analysis?

Exam reliability → Item Difficulty → Item Discrimination
Balancing Item Difficulties on an Exam
Which is the Best Eye Chart?
An Exam Should have a Balance of Item Difficulties
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?

<table>
<thead>
<tr>
<th>Test Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NUMBER OF EXAMINEES</strong></td>
</tr>
</tbody>
</table>

**TOTAL SCORE:**
- **MEAN** 71.21
- **VARIANCE** 72.68
- **STANDARD DEVIATION** 8.53
- **KR-20 RELIABILITY** 0.8393
- **S.E. OF MEASUREMENT** 3.4173
The Anatomy of Question Analysis

**ITEM 42:** DIF = .366, RPB = .430, CRPB = .366 (95% CON = .255, .467)

**RBIS = .551, CRBIS = .469, IRI = .207**

**GROUP**    **N**  **NR**  **NF**  **O**  **1**  **2**  **3**  **4**  
**TOTAL**  257    0    0    0    .163    .280    .191    .366  
**HIGH**    61    0              .049    .148    .131    .672  
**MID**     129    0              .194    .240    .209    .357  
**LOW**     67    0              .209    .478    .209    .104  

**TEST SCORE MEANS**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>28.952</strong></td>
<td><strong>27.264</strong></td>
<td><strong>29.735</strong></td>
<td><strong>35.138</strong></td>
<td></td>
</tr>
</tbody>
</table>

**DISCRIMINATING POWER**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>-.160</td>
<td>-.330</td>
<td>-.078</td>
<td>.568</td>
<td></td>
</tr>
</tbody>
</table>

**STANDARD ERROR OF D.P.**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>.011</td>
<td>.018</td>
<td>.013</td>
<td>.019</td>
<td></td>
</tr>
</tbody>
</table>

**Question 42**

This question meets all major standards.

These statistics indicates the proportion of students among the entire group who chose each answer.

1 = Choice “A” = 16.3%
2 = Choice “B” = 28.0%
3 = Choice “C” = 19.1%
4 = Choice “D” = 36.6%

* indicates keyed response

**DIF:**

The correct answer was selected by 36.6% of all students.

**CRPB:**

This statistic indicates the discriminating power of the question. The higher the number, the greater the question discrimination between high- and low-achieving students. The minimum acceptable in our branch standard is .200.

**Test Score Means:**

These statistics indicate the average score for the group of students that selected each alternative. For example, the average score for students who selected choice A was 28.952 (in this case) out of a total of 50. The average score of the group selecting the keyed response should **always** be higher than the average score for groups selecting each of the distracters.
What guides our use of item analysis?

B. How **difficult** is the item?
   - If machine-scored, check DIF
   - What does this tell you?

   - What is considered difficult, moderate and easy?
What guides our use of item analysis?

C. Is the item positively *discriminating*?
   - If machine-scored, check RBIS & CRPB
   - What does this tell you?
   - What is considered appropriate?

What should you do if its negative?
4. 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

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 the same alternative.
What to look for in your item analysis

<table>
<thead>
<tr>
<th>Difficulty</th>
<th>Discrimination</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greater than .90</td>
<td>Any value</td>
<td>Easy item – desirable to have some of these on the assessment</td>
</tr>
<tr>
<td>Between .50 and .90</td>
<td>Greater than .20</td>
<td>Moderate difficulty and highly discriminating – typical of quality items</td>
</tr>
<tr>
<td>Between .50 and .90</td>
<td>Less than .20</td>
<td>Moderate difficulty, non-discriminating *probably needs adjustment</td>
</tr>
<tr>
<td>Less than .50</td>
<td>Greater than .20</td>
<td>Tough question, highly discriminating *fair to have some of these</td>
</tr>
<tr>
<td>Less than .50</td>
<td>Less than .20</td>
<td>Tough question, does NOT discriminate * toss this one out.</td>
</tr>
</tbody>
</table>
What can you read from this analysis?
**What to do if you don’t machine score?**

**Estimate Item Properties by Hand**

<table>
<thead>
<tr>
<th>Item Difficulty</th>
<th>Item Discrimination</th>
</tr>
</thead>
<tbody>
<tr>
<td># of correct answers</td>
<td>1. Rank-order examinees and select the equal number of them from the highest and lowest scoring group</td>
</tr>
<tr>
<td># of people taking the test</td>
<td>2. Calculate how many in each group got the item right</td>
</tr>
<tr>
<td>( \rho = \frac{\text{# correct}}{\text{total}} )</td>
<td>3. Use the following formula:</td>
</tr>
<tr>
<td></td>
<td>( D = \frac{\text{# correct (up)}}{\text{# correct (low)}} ) # of students in either group</td>
</tr>
</tbody>
</table>
Why is any of this important?

- Does the exam (or items) tell us whether students have learned what we intended for them to learn?

<table>
<thead>
<tr>
<th>Reliability of Assessment</th>
<th>Validity of Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>➢ Consistency of scores</td>
<td>➢ Accuracy</td>
</tr>
<tr>
<td></td>
<td>• Extent to which an assessment method measures what we intend it to measure</td>
</tr>
</tbody>
</table>

http://www.socialresearchmethods.net/kb/relandval.php
References & Time for Questions

- Bloom’s taxonomy:
  http://www.odu.edu/educ/roverbau/Bloom/blooms_taxonomy.htm

- Fairness:

- Item Construction:

- Designing exams: