Series Objectives

After completion of this 4-part series on assessment, the participant should be able to:
- Write high quality multiple choice (MS) and short answer essay questions
- Understand other assessment options (e.g., “hot spot” items, fill in the blank, matching)
- Assess post-test item statistics and overall test reliability
- Evaluate methods of setting a passing standard
- Develop scoring rubrics for essay items
- Use e-technology (e.g., Blackboard, Respondus, Turning Point) to develop, deliver, and evaluate an examination

Writing Good Test Questions

…and then What?

Scored Item Analysis

Overall Measures of Reliability

Setting A Passing Standard

Eric F. Schneider, Pharm.D., BCPS
Associate Dean, Northwest Campus
Associate Professor of Pharmacy Practice
College of Pharmacy

UAMS

After completion of this 4-part series on assessment, the participant should be able to:
- Write high quality multiple choice (MS) and short answer essay questions
- Understand other assessment options (e.g., “hot spot” items, fill in the blank, matching)
- Assess post-test item statistics and overall test reliability
- Evaluate methods of setting a passing standard
- Develop scoring rubrics for essay items
- Use e-technology (e.g., Blackboard, Respondus, Turning Point) to develop, deliver, and evaluate an examination

…from Session 1

- Plan a test based on course goals and objectives (i.e., closing the curricular loop)
- Define the purpose of the exam
  - formative vs summative
  - course level vs program level
  - minimal competency vs assigning a grade
- Determine the level of the items in the assessment tool
  - rote versus application
  - Blooms taxonomy

Planning an Exam:
Close the Curriculum Loop

Competency

Assessment Exercise(s)

Learning Objective

Educational Exercise(s)

Determine Exam Purpose:
Curricular Process

Learning Objective

Summative Exams

Course / Interim Exams

Determine Exam Purpose:
Curricular Process

Planning a Question:
Bloom’s Taxonomy

Knowledge

Comprehension

Application

Analysis

Synthesis

Evaluation

Diagram by Tom Rice, Rice University, MD

UAMS

After completion of this 4-part series on assessment, the participant should be able to:
- Write high quality multiple choice (MS) and short answer essay questions
- Understand other assessment options (e.g., “hot spot” items, fill in the blank, matching)
- Assess post-test item statistics and overall test reliability
- Evaluate methods of setting a passing standard
- Develop scoring rubrics for essay items
- Use e-technology (e.g., Blackboard, Respondus, Turning Point) to develop, deliver, and evaluate an examination

...from Session 1

- Plan a test based on course goals and objectives (i.e., closing the curricular loop)
- Define the purpose of the exam
  - formative vs summative
  - course level vs program level
  - minimal competency vs assigning a grade
- Determine the level of the items in the assessment tool
  - rote versus application
  - Blooms taxonomy

Planning an Exam:
Close the Curriculum Loop

Competency

Assessment Exercise(s)

Learning Objective

Educational Exercise(s)

Determine Exam Purpose:
Curricular Process

Planning a Question:
Bloom’s Taxonomy

Knowledge

Comprehension

Application

Analysis

Synthesis

Evaluation

Diagram by Tom Rice, Rice University, MD

UAMS

After completion of this 4-part series on assessment, the participant should be able to:
- Write high quality multiple choice (MS) and short answer essay questions
- Understand other assessment options (e.g., “hot spot” items, fill in the blank, matching)
- Assess post-test item statistics and overall test reliability
- Evaluate methods of setting a passing standard
- Develop scoring rubrics for essay items
- Use e-technology (e.g., Blackboard, Respondus, Turning Point) to develop, deliver, and evaluate an examination

...from Session 1

- Plan a test based on course goals and objectives (i.e., closing the curricular loop)
- Define the purpose of the exam
  - formative vs summative
  - course level vs program level
  - minimal competency vs assigning a grade
- Determine the level of the items in the assessment tool
  - rote versus application
  - Blooms taxonomy

Planning an Exam:
Close the Curriculum Loop

Competency

Assessment Exercise(s)

Learning Objective

Educational Exercise(s)

Determine Exam Purpose:
Curricular Process

Planning a Question:
Bloom’s Taxonomy

Knowledge

Comprehension

Application

Analysis

Synthesis

Evaluation

Diagram by Tom Rice, Rice University, MD

UAMS
Determine Item Type
- Objectively scored (T/F, MC, short answer)
- Subjectively scored (essay)
- Construct well developed items—using checklists
  - Multiple Choice
  - True-False
  - Short Answer
  - Essay

For subjectively scored items:
- assess with scoring Rubrics
- respond to unanticipated responses

For objectively scored items:
- review item statistics
- respond to faulty items
- Other types of items
  - ordered response, hot spot, matching

Factors Contributing to Maximal Overall Test Validity

What Determines Test Validity & Reliability?
- Item quality
- Test length
- Level of difficulty
- Level of discrimination

How Do We Know that Our Assessment Exercise was Valid and Reliable?
Scored Item Analysis

Scored Item Parameters

- **Difficulty** (the difficulty level of each individual item)
- **Discrimination** (the extent to which success on an item corresponds to success on the whole test)
- **Quality of Options** (how well incorrect answers attracted poorer students and vice versa)


***Difficulty***

- Difficulty \((p^+)\) is traditionally defined as the percentage of students answering the item correctly.
- It is necessary for a test to be composed of items of varying levels of difficulty.
- Goal difficulty target is 0.5 for norm referenced exams.
- Overall target \(p^+\) 0.25-0.75. Questions outside of this range are too hard or too easy and contribute relatively little toward ranking examinees according to their knowledge.
- Norm-referenced testing refers to the process of comparing one test-taker to his or her peers.

***Discrimination***

- Item discrimination indicates the extent to which success on an item corresponds to success on the whole test.
- Any item with negative or zero discrimination undermines the test (reduces validity).

**Discrimination Index** \((r)\) is computed from equal-sized high and low scoring groups on the test.

- Subtract the number of successes by the low group on the item from the number of successes by the high group, and divide this difference by the size of a group.
- Sometimes called a High-Low index.
- The range of this index is +1 to -1.
- Values of \(\geq 0.4\) are high; <0.2 are low.
- Negative values indicate a flawed item.

**Point Biserial** \((rpbi)\) is the Pearson correlation between item score (usually scored right or wrong) and total test score.

- The range of the rpbi is +1 to -1.
- The closer the rpbi is to 1, the more discriminating the item; negative values indicate a flawed item.
- rpbi = .50 or higher are considered excellent; For norm-referenced tests typically in the range of 0.35 to 0.70. Consider re-writing if <0.15.
Item Response Curve

- Graphs the characteristics of a particular item
- Combines effects of difficulty and discrimination
- The total test score is represented on the horizontal axis and the proportion of test takers passing the item within that range of test scores is scaled along the vertical axis.

**Well Discriminating Medium Difficulty Item**


**Well Discriminating High Difficulty Item**


**Well Discriminating Easy Item**


**Poorly Discriminating High Difficulty Item**


**Which Item Performs Well?**

Quality of Answer Options

- Distractors that are not chosen should be replaced or eliminated (not contributing to the test’s ability to discriminate).
- If a single distractor is chosen more than all other options (including the answer), be suspicious about its correctness, especially if:
  - the well performing candidates are choosing it
  - its correlation with the total score is positive

Examples

<table>
<thead>
<tr>
<th>Response Summary</th>
<th>Frequency Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. 0.00%</td>
<td>11 (53.2%)</td>
</tr>
<tr>
<td>b. 0.50%</td>
<td>4 (5.4%)</td>
</tr>
<tr>
<td>c. 100.00%</td>
<td>93 (78.2%)</td>
</tr>
<tr>
<td>d. 100.00%</td>
<td>16 (13.4%)</td>
</tr>
</tbody>
</table>

$p+ = 0.739$
$r = 0.467$

Examples

<table>
<thead>
<tr>
<th>Response Summary</th>
<th>Frequency Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. 0.00%</td>
<td>13 (10.9%)</td>
</tr>
<tr>
<td>b. 100.00%</td>
<td>93 (78.2%)</td>
</tr>
<tr>
<td>c. 0.00%</td>
<td>12 (10.9%)</td>
</tr>
<tr>
<td>d. 0.00%</td>
<td>0 (0%)</td>
</tr>
</tbody>
</table>

$p+ = 0.782$
$r = 0.067$
### Examples

**Response Summary**

<table>
<thead>
<tr>
<th>Answer Value</th>
<th>Frequency Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. 0.00%</td>
<td>10 (8.4%)</td>
</tr>
<tr>
<td>b. 0.00%</td>
<td>51 (42.9%)</td>
</tr>
<tr>
<td>c. 100.00%</td>
<td>47 (39.5%)</td>
</tr>
<tr>
<td>d. 0.00%</td>
<td>11 (9.2%)</td>
</tr>
</tbody>
</table>

- $p^+ = 0.395$
- $r = 0.00$

### Summative Exam 1 Data

<table>
<thead>
<tr>
<th>Item #</th>
<th>$p^+$</th>
<th>$rbpi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.90</td>
<td>0.06</td>
</tr>
<tr>
<td>2</td>
<td>0.90</td>
<td>0.16</td>
</tr>
<tr>
<td>3</td>
<td>0.95</td>
<td>0.14</td>
</tr>
<tr>
<td>4</td>
<td>0.88</td>
<td>0.03</td>
</tr>
<tr>
<td>5</td>
<td>0.67</td>
<td>0.26</td>
</tr>
<tr>
<td>6</td>
<td>0.26</td>
<td>0.05</td>
</tr>
<tr>
<td>7</td>
<td>0.84</td>
<td>0.15</td>
</tr>
<tr>
<td>8</td>
<td>0.93</td>
<td>0.03</td>
</tr>
<tr>
<td>9</td>
<td>0.96</td>
<td>0.14</td>
</tr>
<tr>
<td>10</td>
<td>0.83</td>
<td>0.21</td>
</tr>
<tr>
<td>11</td>
<td>0.94</td>
<td>0.17</td>
</tr>
<tr>
<td>12</td>
<td>0.97</td>
<td>0.03</td>
</tr>
<tr>
<td>13</td>
<td>0.88</td>
<td>0.11</td>
</tr>
<tr>
<td>14</td>
<td>0.91</td>
<td>0.16</td>
</tr>
<tr>
<td>15</td>
<td>0.88</td>
<td>0.36</td>
</tr>
</tbody>
</table>

### Overall Reliability of an Examination (Internal Consistency)

- Summative Exam
  - Tests college-wide competencies
  - Tests *global* knowledge / important concepts
  - Intent is not to discriminate among high performing candidates (i.e., easy items are less worrisome)
  - Separates minimally competent from NOT competent
- Setpoints
  - Difficulty: between 0.25 and 0.95
  - Discrimination: above 0.1
Tests of internal consistency measure reliability within the instrument (i.e., consistency among the questions). In general, the more items, the more reliable the test. A reliability of 0.8 indicates 80% consistency in the scores produced by the test. A given exam should have a minimum reliability score of 0.7 to be considered reliable (0.9 for achievement tests).

Methods of Assessing Internal Consistency

Kuder-Richardson Formulas
- Items on the instrument dichotomously scored
- All items are compared with each other
- Kuder-Richardson Formula 20 (KR-20)
  - Does NOT assume all questions equally difficult
- Kuder-Richardson Formula 21 (KR-21)
  - Does assume all questions equally difficult
  - Easier to perform, but less precise

Cronbach’s Alpha
- Flexible and can be used with test formats that have more than one correct answer (continuous measures)
- Used when items on a test are not scored as right vs wrong
- Cronbach’s Alpha = KR-20 when items are dichotomous

Summative Exam 1 Results

Spring 2010

| Mean for Exam | 76.4 |
| Standard Deviation for Exam | 7.48 |
| Split-Half (odd-even) Correlation | 0.59 |
| Spearman-Brown Prophecy | 0.74 |
| KR21 | 0.68 |
| KR20 | 0.75 |
| Cronbach's Alpha | 0.75 |

The Big Picture
### Assessing Reliability

- A high reliability coefficient tells whether the items in the assessment are strongly related with regard to student performance.
- A high reliability coefficient is no guarantee that the assessment is valid.
- Consider reworking items with:
  - discrimination < 0.15
  - unused distractors

### Methods for Setting A Passing Standard

#### Angoff Method

- A standard setting process is based on the hypothetical group of **minimally acceptable** students, with estimates of the percent of **minimally acceptable** candidates expected to answer the item correctly.
- Uses subject matter experts (SME’s) as judges.
- Each SME predicts a mean overall score for the minimally competent candidate.
- Passing score is the average of the SME scores.

#### Modified Angoff Method

- A standard setting process is based on the hypothetical group of **minimally acceptable** students, with estimates of:
  1. how critical the knowledge is for a **minimally acceptable** candidate.
  2. how difficult the item would be to the **minimally acceptable** candidate.
  3. the percent of **minimally acceptable** candidates expected to answer the item correctly.
- Uses subject matter experts (SME’s) as judges.

#### Angoff Yes/No Method

- The items a hypothetical group of **minimally acceptable** students should answer correctly are assigned a 1 = yes, and the items they should not answer are assigned a 0 = no. The pass point is then calculated by averaging the scores.
- Uses subject matter experts (SME’s) as judges.

#### Nedelsky Method

- Judges independently identify the answer option(s) that a borderline candidate would be able to recognize as implausible.
- The number of remaining options determines the probability that the candidate will answer correctly the item: 1 plausible response = 100%, 2 = 50%, 3 = 33%, 4 = 25%, and 5 = 20% probability of a correct response.
- The pass point is the sum of the means of the probability for success.
### Bookmark Method
- Item difficulties are determined from data from an actual exam administration.
- Items are arranged by item difficulty from least to most difficult.
- Judges select the most difficult item a borderline candidate would be likely to answer correctly and a "bookmark" is placed at that location.
- The pass point is set by summing the number of items up to the bookmark for each judge and averaging across judges.


### Borderline Group Method
- A description is prepared for each performance category (clearly inadequate, borderline, clearly competent).
- SMEs are asked to submit a list of participants whose performance on the test should be close to the performance standard (borderline).
- The median test score is used as the cut score.


---

Where do we go from here?

---

**...For Next Time (7/28/10)**

- Developing Grading Rubrics (Eric Schneider)
- Computerized Testing (Martha Carle)
  - eTesting in Blackboard
  - Setting up a Respondus template
  - Using Lock Down Browser.
  - Using your questions in Turning Point
  - Academic Integrity.
  
  The second hour will be hands on lab time.
  COPH 1230/1250