Day 2 Session 10:45 to 12:00 noon

Developing Quality Assessment Items
Constructing good quality selected response items
In this session we will

- review the requirements for developing effective large scale assessments
- define selected and constructed response items
- match item formats with learning outcomes
- review the Characteristics for assessment methods
- describe the life cycle of an item
Requirements for developing effective large scale assessments
Developing effective large scale assessments

- Make sure there is a **clear purpose** for the assessment.

- Make sure that the **learning outcomes** are clearly **articulated in the assessment framework**.

- Make sure that the items, tasks, etc. are **clearly linked** to the **learning outcomes**.

- Make sure that the **items** are of **good quality**.

- Make sure that the **results** are **reported effectively**.
Define selected and constructed response items
Selected and constructed response items

- All test items require students to make a **selected-response** or a **constructed-response**.

- **Selected-response** type items require students to choose an answer from a set of two or more options (e.g. multiple choice items and matching items).

- **Constructed-response** items require students to supply their own responses (short answer, essay, performance assessments and other instruments).
Match item formats with learning outcomes
## Matching selected and constructed response to outcomes

<table>
<thead>
<tr>
<th>Selected Response</th>
<th>Constructed Response</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Demonstrate knowledge and understanding</strong></td>
<td><strong>Selected Response</strong></td>
</tr>
<tr>
<td>Multiple Choice Matching</td>
<td>Essay, Short Answer</td>
</tr>
<tr>
<td>Generally used for assessing knowledge</td>
<td>Can be used to assess knowledge and understanding</td>
</tr>
<tr>
<td><strong>Solve problems and reason</strong></td>
<td>Can be used to assess problem solving and reasoning – difficult to write</td>
</tr>
<tr>
<td>Can be used to monitor students as they solve problems; and, then infer the ability to reason and solve problems</td>
<td>Can ask students to describe out loud what they are doing</td>
</tr>
<tr>
<td><strong>Demonstrate performance skills</strong></td>
<td>Generally not used to assess the actual skill, but can be used to assess the knowledge component related to the skill</td>
</tr>
<tr>
<td>Can use interview schedule to assess the communication parts of the performance and knowledge part of the performance</td>
<td></td>
</tr>
<tr>
<td><strong>Create products</strong></td>
<td>Can assess knowledge part of the ability to create products, but cannot use these to assess the products themselves</td>
</tr>
<tr>
<td>Can ask students about the procedure they are using to create the product, but not ask about the product itself</td>
<td></td>
</tr>
</tbody>
</table>
Characteristics for assessment methods
Assessment Methods

- Curriculum relevance
- Fairness
- Instructional utility
- Reliability/comparability
- Practical convenience
Curriculum Relevance

Selected and constructed response items

- assess the particular outcome that they are supposed to measure (validity);
- be “in harmony” with the requirements of the curriculum;
- be constructed so that students at different locations along the continuum can demonstrate where they are located
- (generally) utilise a range of different tasks/item formats: What is the most appropriate way to find out what it is a student knows and can do relative to this outcome?
- be written to engage the students.
Instructional Utility

- Selected and constructed items and tasks can provide useful instructional material if they are constructed properly.
- Selected response items can be derived from constructed response items.
Reliability/Comparability

• Selected response items are usually objectively marked (no subjectivity ~ hence greater reliability)
• More items assessing a range of outcomes increases coverage (sampling) of the construct, greater reliability
• Items that are too easy or too hard give low reliability ~ spread items over a range of difficulties
• Generally the more selected response items in the test, the greater the reliability
• The more objective, the greater the comparability.
Practical Convenience

- Trade-off between validity and reliability ~ lots of items (greater reliability ~ less validity)
- Trade-off between validity and practical convenience (e.g. assessing one-on-one) is more valid but not particularly practical.
Items must be “fair” (valid). They must

- present a single, clearly formulated problem in the stem;
- have one correct or clearly best answer;
- have distractors that are plausible and attractive and are linked to “incorrect” strategies;
- ensure that one option does not stand out from the rest, for such reasons as, grammatical consistency, similarity in wording, more explicit language/length;
Fairness - Selected response items

Items must be “fair” (valid). They must

• not contain specific determiners (always, all, never, only, none);
• not advantage one group over any other group (e.g. gender, culture);
• have clear directions;
• have appropriate level of reading vocabulary and sentence structure; and
• be editorially correct and the language must be appropriate.
Quick Quiz (True or False or Don’t Know [Need More Information])

A GOOD ITEM

1. Measures a specific learning objective.
2. Contains subject matter and vocabulary that is above the student’s grade-level.
3. Has only one correct answer or clearly best answer.
4. Assesses trivial or obscure subject matter.
5. Free from grammatical clueing.
6. Is free of negative wording such as “not” or “none of the above.”
Quick Quiz (True or False or Don’t Know [Need More Information])

A GOOD ITEM

7. Assesses more than one concept.

8. Contains options that are opposite of one another.

9. Contains distractors that assess common errors or misconceptions.
Quick Quiz (True or False or Don’t Know [Need More Information])

1. Distractors should be parallel in content, structure, and length.

2. Cognitive level refers to the difficulty level of an item.

3. Items should be written so that the content in the item is accessible to the widest range of students.

4. Parallelism refers to when a student from different ethnic, sex, or cultural groups perform differently on an item.

5. Controversial items are often assessed with the multiple-choice format.
Quick Quiz (True or False or Don’t Know [Need More Information])

6. Item fairness means that the item assesses all students at the appropriate age and enrolled grade-level.

7. An example of bias is presenting a culturally stereotypical situation in the item.
Fairness – Constructed response items (Short Answer Items)

Short answer Items must be “fair” (valid). They must

- align with primarily one learning outcome.
- have only one correct answer.
- have a short response (if possible).
- generally be presented as a direct question fin preference to the complete the sentence format: This is the way that it is generally asked in class.
- not contain any unintentional clues.
- have a marking key and it is applied consistently.
- not be written in a manner that advantage or disadvantage any particular group of students
Essay and performance items must be “fair” (valid). They must

- have the question written in simple, clear and direct language: making sure there is enough information to ensure that the students know exactly what is required in the task.
- enough time for them to complete the essay or performance item.
- be written at the cognitive level indicated by the verb in the standard or learning outcome.
Fairness – Constructed performance tasks (Essay Items)

Performance tasks must be “fair” (valid). They must

- be used to measure qualities such as analysis and research skills with relevant evidence
- allow for multiple approaches
- allow for demonstration of important knowledge and skills, including 21st Century Skills such as critical analysis and synthesis that are presented in a variety of formats and media.
- be the most direct assessment of the learning outcome; try and make sure it encompasses more than one learning objective; select performance tasks that assess learning objectives that are teachable.
- have clear instructions that describe the purpose of the task and how the students are to respond.
- clear marking rubrics (essentially the same as essay items): rating scales and checklists are typically used as rubrics for performance items.
Life cycle of an item/task
Select an outcome/s

Reading passages to Review Committee

Surviving passages used as a basis for item writers to write items

Draft items to Review Committee

Items either retained; retained for modification; or eliminated

Surviving items brought back to item developers for modification

Surviving items used to build pilot test forms

Revised items brought back to Review Committee

Pilot tests administered to students

Pilot test results analysed

Item Statistics to Review Committee

To final tests