Van Blerkom Chapter 7: Essay Items

 Essay items enable one (a) to measure any skill level and (b) to measure one’s ability to communicate. Both advantages and limitations of essay items are discussed below.

1. Advantages and Limitations of Essay Items

1a. Advantages of Essay Items Relative to Other Item Formats

* Essay items usually provide a better measure of behavior specified by performance objectives. Essay items enable one to more easily incorporate higher-level cognitive skills such as synthesis and evaluation, while most other formats can only measure these skills indirectly.
* Essay items enable one to demonstrate communication via writing. While this is true, one must remember that scores assigned to essay responses should reflect a distinction between writing/communication skill and the content focus found in essay responses.
* Since essay items require students to supply responses, they tend to be reduce the effect of guessing upon reliability and validity. However, as will be noted below, the difficulty in assigning grades consistently to essay items reduces reliability, and since one cannot easily often increase the number of essay items on a test, reliability lost due to inconsistent scoring is unlikely to be regained.

2b. Limitations of Essay Items

* Since fewer essay items can be included within a given test due to time constraints, essay items do not allow for a large sampling of content like short-answer, alternate-response, and multiple-choice formats. This drawback will most likely affect content-related evidence of validity for a classroom test.
* Scoring reliability of essay items is typically much less than for alternative item formats due to the subjective nature of essays.
* Essay items require more time to score than do alternative item formats.

2. Qualities Desired in Essay Items

 (a) Desirability of Using Brief-Response (or Restricted-Response) Versus Extended-Response Questions

 A **brief-response essay item** is one that requires no more than 10 minutes for students to both read the item and provide a response. If more than 10 minutes is required, then the item is referred to as an **extended-response essay**, or simply an **essay**. Brief-response essay items enable one to include a better sampling of content than do essay items. Note also that because extended-response essay items often require considerable time to complete, they are not recommended for classroom tests due to the low level of reliability commonly associated with these types of items, but they can be used successful for non-classroom, or out-of-classroom testing situations.

(b) Criteria for Evaluating Essay Items

* **Does the item measure a specific skill?** Following a table of specifications with help ensure that the essay item relates well to desired content and skills.
* **Is the item written at a level below the targeted students’ reading ability?** Writing items at a lower level helps prevent confounding reading ability with achievement.
* **Will most all students answer this item in 10 minutes or less**? If not, then perhaps this item should not be included on a classroom test as it will prevent a more detailed sampling of content and will likely reduce the reliability of the test.
* **Does the scoring plan, developed when the item is written, enable knowledgeable people to assign consistent scores to the essay?** The goal of such scoring plans is to allow for people knowledgeable in the given content to assign with consistency scores to essays, i.e., allow for high scorer/rater agreement. Aspects of an essay scoring plan are listed below.

|  |
| --- |
| Good scoring plans may be developed using the following three criteria:* Clearly specify the total number of points an item is worth, and these points should be proportional to the relative importance of the skill the essay item is assessing.
* Each attribute to be assessed must be clearly specified in the scoring plan. Typically, poor grammar and irrelevant information should not be included in scoring plans for essay responses unless the content area focuses upon grammar.
* Lastly, how points are to be awarded for each attribute scored must be established. The scoring plan must be detailed so readers will know when to award or subtract a point, and when to give partial credit.
 |

* **Will content experts agree the scoring plan outlines the correct response or correct procedure for this item?**
* **Will knowledgeable students be able to determine the scoring plan for an item when they see the item on a test?** Ideally, knowledgeable students should be able to point to important components of any essay items, and should be able to describe, approximately, the weight assigned to each component.

 (c) Assessing Mathematical Skills with Essay Items

 Asking students to provide algebraic expressions or solutions to complex mathematical problems represent essay items. The primary difference between verbal and mathematical essay items is typically the length of the response, with mathematical responses usually shorter.

3. Scoring Students’ Responses to Essay Items

3a. Analytical Versus Holistic Scoring

**Analytical scoring** results when one uses a specific, well-defined scoring plan for assigning scores to essay responses. However, it may be difficult to develop well defined scoring plans for complex or broad essay items. For such items, **holistic scoring** may be used. With holistic scoring, one typically reads an essay response and then classifies the response, as a whole, into one of three groups, say high, middle, and low. Once these three groups are formed, further classifications are developed based upon a re-reading of the essays. One strategy for further re-classification is to create as many classifications as there are points associated with the essay item. This, of course, becomes problematic when many points are assigned to an essay item.

Holistic scoring is generally quicker than analytical scoring but is limited in that it does not provide a clear indication to students as to why certain points were assigned to his or her response, and does not indicate problem areas as well as analytical scoring. Because of these practices, holistic scoring is less reliable than analytical scoring. One approach to increasing reliability with holistic scoring is to develop a model answer. This model answer can also be distributed to students after the examination as a form of feedback.

3b. Reading Responses to Multiple Versus Single Items

 When evaluating essay items, it is best to evaluate all students’ responses to a single item before moving to another item. This helps ensure reliability by better enabling one to assign scores consistently for similar mistakes. To assist in scoring, one should sort responses into piles of similar mistakes before assigning points.

3c. Reading Students’ Papers in a Consistent Versus Variable Order

 Research has shown that bias can be introduced into scoring when one reads students’ responses in the same order for each essay item. If one follows the procedure outlined in 3b above (i.e., sorting essays into piles of common mistakes), students’ papers will be rearranged throughout the scoring process.

3d. Concealing the Identity of Students

 Also, to further prevent bias, student’s identities should be concealed if possible. One method for doing this is to ask students to place only an identification number, rather than their names, on their work.

3e. Use Multiple Readers

 With both holistic and analytical scoring, it is beneficial to use multiple readers to assign scores. If one takes the average score assigned across multiple readers, reliability will be increased. Highly variable scores for a given essay response should not be obtained, if so, then one should examine the given essay in detail to determine why high variability in scoring exists.

3f. Use of Diversity of Responses as an Indicator of Item Ambiguity

 If one obtains highly diverse responses to a given essay item, then the item is unclear and ambiguous. Diverse responses are a clear indication that a poorly constructed item exists. Such diversity lowers validity since students are interpreting the item in different ways.

Self-Test: Essay Items

For items 1 through 5, indicate (yes or no) whether each statement describes an advantage of essay format.

1. Allows testing a student’s ability to communicate ideas in writing.

2. One of the better item formats for measuring many course objectives in a single test.

3. Can measure more types of skills than the multiple-choice format.

4. Results in more reliable test scores than other item formats.

5. Requires more time to score than other item formats.

Items 6 through 13 propose characteristics one might see in essay items. Indicate whether each characteristic is

(A) Desirable

(B) Undesirable

6. Students should require between 15 and 20 minutes to answer each essay item.

7. The amount of time required to answer an item should determine the number of points associated with the item.

8. Students should be told how many points are associated with each item.

9. More points should be assigned to essay than multiple-choice items because more time is required to answer essay items.

10. Points should be subtracted for spelling and grammar errors regardless of whether these attributes are addressed by the scoring plan.

11. Points should be subtracted for factual errors within answers superfluous to what was asked for in the essay item.

12. The scoring plan should give readers considerable flexibility in determining what constitutes a good response.

13. An essay item should be worded so students knowledgeable in the content being tested can describe approximately how points will be awarded to answers.

Items 14 through 17 propose techniques one might use when scoring essay tests. Indicate whether each technique is

(A) Desirable

(B) Undesirable

14. Particularly with holistic scoring, have more than one reader score each response.

15. Identify which student wrote the response being graded to better understand what the student meant to say.

16. Read one student’s responses to all items before scoring another student’s answers.

17. Shuffle students’ papers into a different order before scoring each item.