Essentials of language testing for Malaysian teachers

Arshad Abd Samad, Universiti Putra Malaysia
CHAPTER 3

TEST FORMATS IN LANGUAGE TESTING

In previous chapters, we have seen that there are various test formats in language testing. In this chapter, we will explore two test formats that are very commonly used in language testing as well as in the testing of other skills and knowledge. The two test formats are the multiple choice question format and the essay type format. We will discuss the two formats according to the following outline:

- The Multiple Choice Format
  - The structure of the multiple choice format test.
  - Use of the multiple choice format.
  - Constructing the multiple choice item.
  - Criticism of the MCQ format.

- The Essay Format
  - Different forms of the essay test.
  - Advantages of the essay format.
  - Shortcomings of the essay format
  - Constructing essay questions
  - Approaches to scoring essays

- Applications

The Multiple Choice Format

The multiple choice format is perhaps the most common test format to many of us. It is also commonly referred to as an objective test as there is seen to be “objectivity” in grading the test. In this section, we will examine the multiple choice format with respect to its structure, use, and construction.

The structure of the multiple choice format test. The test consists of several basic but important components which are the rubrics, the stem, as well as the distractors and the key, which together make up the alternatives or options. These components are illustrated in the following diagram:

![Diagram of multiple choice question structure](image)

Figure 5. The structure of the multiple choice question
In the diagramme above, the major components of the multiple choice format are clearly laid out. It should be noted that the number of alternatives in a multiple choice item usually ranges from between three to five.

**Using the multiple choice format.** When would you use multiple choice format tests? There are a number of situations in which a multiple choice format test may be useful and appropriate. Ory (n.d.) outlines some of these situations as follows:

- When there is a large number of students taking the test
- When you wish to reuse the questions in the test
- When you have to provide the grades quickly
- When highly reliable test scores must be obtained as efficiently as possible
- When impartiality of evaluation, fairness and freedom from possible test scoring influences such as fatigue are essential
- When you are more confident of your ability to construct valid objective test items clearly than of your ability to judge essay test answers fairly
- When you want to sample a wide range of content
- When you are especially interested in measuring particular learning objectives such as comprehension, recognition, and recall
- When you want specific information especially for diagnostic feedback

It should be noted that these situations reflect the advantages of using the multiple choice question format. These advantages include:

1. the ability to create a test item bank
2. quick grading
3. high reliability
4. objective grading
5. wide coverage of content
6. precision in providing information regarding specific skills and abilities

**Constructing the multiple choice item.** While the multiple choice question format may have its benefits, the construction of the MCQ test is not easy. There are many issues that need to be considered in constructing the test including how we come up with distractors in the item as well as how we can construct good test items.

**Selecting distractors**

It is important that distractors in a multiple choice test format perform their task of distracting students from the correct answer. If they do not, then they might as well not be included as a distractor. Good distractors are not always easy to create. Thankfully, however, there are a number of sources from which we can obtain reasonable distractors. These sources include:

- **Previous incorrect responses.** If we have information on the kinds of errors students make, we can use such information to create distractors for an MCQ item. Information such as this may be obtained from classroom exercises, research findings, or even a pilot test involving fill in the blank type questions.

- **Incorrect form, meaning, style or register.** A distractor may be incorrect because of its grammatical form, the wrong meaning it conveys, or its inappropriate style or register. In the following item, for example, alternative B cannot be the correct response because it is inappropriate in...
the social context of a conversation between father and son. We would expect a son to be more polite when talking to his father.

Father: Son, too much sports can be harmful.
Son: …(to disagree)……

A. But, dad, I think it is good exercise
B. No, that’s definitely not true
C. I strongly disagree with your opinion
D. Well, that’s true in many cases

Alternative C is not as good a choice as alternative A because it is rather formal and therefore not consistent with the somewhat casual tone of the conversation. Finally, alternative D is not correct simply because it does not express a disagreement.

Translations from the native language. Many students tend to rely heavily on their native language in interpreting a task and in responding to a question. Often, these translations are inaccurate. Distractors that are formed based on inaccurate translation can be appropriate in multiple choice format tests. Take the following item as an example:

My mother did not come because she is _____

A. pain      B. sick     C. injury     D. ache

Alternative A is chosen as a distractor because it is an inaccurate translation of the Malay word ‘sakit’. The same can be said about alternative D. The word ‘sakit’ can be translated into pain, sick, as well as ache, but it is clear that in this sentence, the only acceptable answer is the word sick.

Variations of the key. Distractors can also be in the form of variations of the key. This may include a homonym, an antonym or even a word which has some graphemic resemblance. For example, if the key is the word take, distractors could be the antonym give, the near synonym get or a near homograph such as teak.

Contextual relevance. The stem as well as other forms of stimuli such as a comprehension text creates a context that is a source of many appropriate distractors. For example, test candidates may be given a text about schools and asked how most students in the text go to school. If the answer is by bus, contextually relevant responses include distractors such as: by car, by taxi and by bicycle which are can all be used as effective distractors.

Constructing Good Test Items.

In addition to the difficulty of finding good distractors, test items may also become poor measures of a student’s ability because they are poorly written and provide unintentional cues to the students. What this does is contribute to the inaccuracy of our measurement. Instead of measuring proficiency in the language, part of the score the student obtains is because of his or her test wise-ness or his ability to notice these unintentional cues. We therefore need to be very careful of the unintentional cues that may creep into our test items. A second issue is that student reaction to how a test item is written may actually lead us to measure that reaction rather than the behaviour or ability we are interested in. Both these issues will be discussed in the following sections:

Unintentional cues in the item.

If we are not careful, our test may be contaminated by unintentional cues that lead students toward a particular answer. Unintentional cues in an MCQ test include:
1. inconsistent length of alternatives
2. inconsistent categories of alternatives
3. convergence cues
4. nonsense distractors
5. common knowledge response

Each of these are described and briefly discussed as follows:

**Inconsistent length of alternatives.** Read the following multiple choice type item and try to find the cue that is unintentionally provided in the item.

In the story, he was unhappy because _____

A. it rained
B. his dog died
C. he was scolded
D. his father had to send his dog back to the pet shop

The way this item is written provides undue attention to the last alternative. As we are interested in measuring the student’s true language ability – comprehension, in this case – it is unfair to draw so much of the students’ attention to a single alternative as in this example. If indeed any of the students select this alternative, it will be difficult to determine whether this was because he was unduly attracted to its obviously different length or whether he or she genuinely believes it to be the answer. It is bad enough that option D does this but it is even worse if it is the correct answer. If this is the case, then we will not be able to tell whether the selection of this alternative was because the student actually knew the answer or because he was at a total loss and chose the alternative simply because it stood out.

**Inconsistent categories of alternatives.** Look at the following item and consider what cue it provides to the test taker.

A. spoke B. screamed C. smiled D. talkative

In the list of distractors above, all the options except D. talkative are verbs. This gives undue attention to that option and increases the likelihood that it will be selected regardless of whether or not it is the key. Another possibility is that the option will be immediately rejected because it is so different from the rest of the options.

**Convergence cues.** The cue that is provided in the following item is a little less obvious and may occur more often in our items without detection. I have excluded the stem for this item in order to better illustrate how it may work.

A. tea B. coffee C. lunch D. dinner

Which of the following would you choose? Without the stem, it is highly likely that many would select A. tea because the two options tea and coffee refer to hot drinks and are immediately related. Between the two options, tea would probably be selected instead of coffee because like the two other options, it describes a meal that many of us take in a day. Therefore, these cues seem to converge on the option A. tea as the most likely answer.

**Nonsense distractor.** Another kind of problem is the nonsense distractor. Unless the students are actually experiencing problems in the language and the teacher intentionally inserts nonsense words into the item as a distractor in order to determine whether or not the student has such problems, the nonsense distractor cue should be avoided. Look at the following alternatives:
A. had gone  B. had went  C. has been going  D. have gone

One of the options above does not exist in the English language. Which one is it? *Had went*, option B, is not a valid construction in the English language and so should be immediately deleted from the list of alternatives. However, once again, it should be made clear that nonsense distractors may be acceptable if the intention is to confirm whether a student has a problem with a particular structure. For example, often Malaysian students tend to leave out the auxiliary *do* in sentences and say or write “Why you say that?” instead of “Why did you say that?” A teacher, therefore, may justifiably decide to include the ungrammatical or nonsense distractor “Why you say that?” in order to confirm whether or not students have problems with the structure.

**Common knowledge response.** Another problem with select type test formats is that the student may select an answer based on common knowledge rather than the skill that we intend to assess. For example, let us consider the following test item:

We know from the reading passage that Hang Tuah is from

A. Johor  B. Kelantan  C. Kedah  D. Melaka

While it is true that the question may have come from a reading passage, the answer is obvious even to those who have not read and understood the passage.

**Reaction to the Item.**

The test context can already be quite taxing and anxiety causing. Test writers should avoid adding to this anxiety. Instead, they should consciously attempt to help students by reducing anxiety, time wastage and all other unnecessary reactions in the test. Some of these problems in writing test items related to this issue are discussed in this section. They include problems such as redundancy, trick questions, and improper sequencing of alternatives.

**Redundancy.** Some items are written in such a way that the students need to read unnecessary information. In a test where time is often a premium, we should avoid writing items in such a way. The following is an example of a test item which includes redundant language.

The boy bought the magazine

A. because he wanted to read it  B. because he saw his face on it  C. because he was told to do so  D. because he needed some change

Each alternative begins with the phrase *because he*. It would have been much more helpful to the students if the phrase had been included in the stem so they would not have to spend time reading it over and over again.
Trick Questions. There are also times when test items provide unnecessary anxiety to the students. In the following example, the item is written in such a way that the student needs to read it over in order to understand what it actually requires.

When is it not appropriate for a student not to be absent from class?

A. When he is sick   C. When class is in session
B. When he is young   D. When the teacher is absent

The use of multiple negatives in the stem – *not appropriate, not to be absent* – make the stem somewhat confusing. Since our intention in a test is not to confuse the student but rather to measure his true ability in language, this sort of item should be avoided.

Improper sequencing. Disorganised sequencing of alternatives also contribute to unnecessary additional cognitive load for students. If the alternatives are days in a week or months in a year, then they should be arranged in chronological order. Similarly, for testing reading comprehension, the alternatives that are taken from the passage should be arranged according to the order that they appear in the passage.

Concerns and criticism of the MCQ format. Although the multiple choice format has been the mainstay of many examinations, it is not immune to criticism. Several of the criticisms that have been leveled against the MCQ format test relate to the difficulty of constructing good items for the test, the limitations with respect to the kinds of questions that can be asked, and the negative effect that it may have on teaching and learning.

Hughes (1989) lists several other problems with the test format which are as follows:

1. The technique tests only recognition knowledge and recall of facts
2. Guessing may have a considerable but unknowable effect on test scores
3. The technique severely restricts what can be tested to only lower order skills.
4. It is very difficult to write successful items, due especially to the difficulty of finding good distractors.
5. Backwash can be harmful – for both teaching and learning as preparing for a multiple choice format test is not reflective of good language teaching and learning practice.
6. Cheating may be facilitated
7. It places a high degree of dependence on the student’s reading ability and instructor’s writing ability
8. It is time consuming to construct

Hughes (1989: 3) goes on to say that “good multiple choice items are notoriously difficult to write. A great deal of time and effort has to go into their construction. Too many multiple choice tests are written where such care and attention is not given (and indeed may not be possible). The result is a set of poor items that cannot possibly provide accurate measurements.”

Others have echoed similar concerns about the multiple choice format, including a serious comment by Oller (1979) who states that it is intrinsically in opposition to the interests of instruction. He dismisses the reliability and ease of administration of the multiple choice format as providing minimal benefits compared to the detrimental effects in instruction and the difficulty of preparation.

The Essay Format

Like the multiple choice test format, the essay is a common technique in testing. However, there are also some critical differences. While the multiple choice question format involves selecting the answer, in essays students are expected to supply the answer. There are also different issues with the
essay, especially with respect to grading, compared to the multiple choice question format. In this section, these issues will be discussed.

**Different forms of the essay test.** The essay can take many different forms ranging from the short response answer to the extended response answer essay. In language testing we can also consider the structured essay that involves directed writing and the free expression or continuous response essay. In the Sijil Pelajaran Malaysia, both these types of essay questions are found in Paper 2 of the English Language paper. The following essay questions, taken from the 2001 examination, are examples of a structured or directed essay and a free writing essay or continuous essay respectively.

---

**Section A: Directed Writing**

**Question 1**

You find that many of your friends are under a lot of pressure. They spend most of their time studying and appear to suffer from stress. Write an **article** for your school newsletter on how to cope with stress. Include the following points:

- regular exercise
- hobbies
- plan a time-table
- work in groups
- balanced meals
- enough sleep
- study skills
- make a check-list
- seek advice
- games
- talk to your family
- listen to music

You should use all the points given above and add **two new points** of your own.

When writing the **article**, you should remember:

- to give a title
- that your readers are mainly students
- to include all the points given
- to add **two new points** of your own
- to write in paragraphs

[30 marks]

---

The directed writing test item above provides structure to the task. Students are expected to respond by using the prompts that are provided in the item.
Section C: Continuous Writing

Write a composition of about **350 words** on **one** of the following topics:

(a) How do you think the media has influenced you and your family?

(b) Some people think that the legal age of driving should be increased from 17 to 19. What do you think?

(c) Describe how you spent a free day when the weather was very bad.

(d) Large families make happy families.

(e) Write a story of an old man returning to his home he left many years ago.

[40 marks]

Unlike the directed writing task, the continuous writing test item provides little structure other than the question itself. Students are expected to draw upon their experience and past knowledge as well as knowledge of writing conventions and organization in order to complete the task.

**Advantages of the essay format.** The essay test format provides several advantages compared to the multiple choice test format. Some of these advantages as mentioned by Kubiszyn and Borich (2000: 18) are:

1. It can assess higher order skills. Unlike the multiple choice test format which is often limited to assessing low order skills, the essay places a premium on the ability to analyse, synthesise and evaluate through topics that require students to express their opinions or argue a point.

2. Emphasises communication skills. This is especially important when we consider that communication skills is an important aspect of social relations.

3. Eliminates guessing. The multiple choice question format is notorious for allowing students to guess. In the essay format, however, guessing is unlikely to occur.

4. Relatively easy to construct. An essay question can be constructed within minutes compared to other test formats which can even take days to construct.

It is quite clear that these advantages have placed the essay in good stead and made it one of the most common test formats.

**Shortcomings of the essay.** While the essay clearly provides some advantages, there are also disadvantages of using the essay format in tests and assessment. In discussing these shortcomings, Kubiszyn and Borich (2000: 119) point out that by using essays, the teacher has limited the content that can be tested. A single essay question which may take about at least half an hour to complete can only test a limited amount of content in comparison to the number of multiple choice type questions in the same time period. Secondly, the essay test is prone to bluffing by the students. Many teachers have come across essays in which students are not able to say much but keep on writing to pad their answers with the hope that the teachers will become impressed by the length of the essay. A third shortcoming pointed out by Kubiszyn and Borich is that essays tend to be less reliable than many other test formats. This is especially so if reliability refers to how consistently graders mark essays. Different graders, for example, may have different impressions of a single essay and consequently
assign it different values. There are also many factors that can affect grading including preconception of students, time of day, as well as neatness and handwriting. For this reason, it is recommended that we control the situation by providing clear guidelines to the students on how and what to write. In other words, in order to control for reliability, it may be better to use the structured essay rather than the free or continuous writing essay. Finally, and in relation to the third point, the essay is also difficult to score. There are many criteria that can be used to grade the essay and graders will have to decide which scoring approach to use. Additionally, unlike the multiple choice test format which is quick to score, the essay will occupy a huge chunk of your time when you have to grade them.

Constructing Essay Questions. Effective essay questions will take advantage of the strengths of the essay format described earlier. Essay questions should therefore focus on higher order thinking skills rather than factual type questions as facts can be more effectively assessed through select type test formats. Similarly, essay questions should take communicative effectiveness into consideration. In order to ensure that students communicate when they write their essay, they must be informed of the importance to pay attention to discourse and rhetorical elements as well as style in their essays. The following are suggestions on writing good essay questions, some of which are gleaned from various sources:

1. Preview the essay question by writing a model answer or response and identifying the characteristics of the model answer that will earn credit.
2. Review the topic and prompt to ensure that it is clear and not ambiguous.
3. Consider the use of several short essay questions rather than one long essay in order to cover more content or allow the students to demonstrate language ability in different contexts or situations or for different communicative purposes.
4. The topic should be meaningful and interesting to the students and also consider their personal experience (Cohen, 1994: 311)
5. Indicate the maximum number of points to be awarded for the essay question.
6. Indicate what the student should do. For example, are they expected to compare and contrast, to simply describe a process or to give an opinion.
7. Suggest the amount of time the student should spend on writing the essay.
8. Indicate the audience to be addressed (Cohen, 1994: 310)

Suggestions number 1 to 4 concern the development of essay questions while suggestions 5 to 8 are specific suggestions that appear in the test prompts.

Approaches to Scoring Essays. As we have seen earlier, scoring an essay is not easy as graders can be easily swayed by many factors. Scoring, therefore, remains one of the major issues in grading essays. There are generally three major approaches to scoring essays which are the holistic scoring method, the analytical scoring method, and the objective scoring method.

Holistic Scoring

In holistic scoring, the reader reacts to the students’ compositions as a whole and a single score is awarded to the writing. Normally this score is on a scale of 1 to 4, or 1 to 6, or even 1 to 10. (Bailey, 1998: 187). Each score on the scale will be accompanied with general descriptors of ability. The following is an example of a holistic scoring scheme based on a 6 point scale.

<table>
<thead>
<tr>
<th>Rating</th>
<th>Criteria</th>
</tr>
</thead>
</table>
| 5      | * Vocabulary is precise, varied, and vivid
* Organization is appropriate to writing assignment and contains clear introduction, development of ideas, and conclusion
* Transition from one idea to another is smooth and provides reader with clear understanding that topic is changing
* Meaning is conveyed effectively
* A few mechanical errors may be present but do not disrupt communication
* Shows a clear understanding of writing and topic development |
<table>
<thead>
<tr>
<th>Band</th>
<th>Descriptors</th>
</tr>
</thead>
</table>
| 4    | * Vocabulary is adequate for grade level  
     * Events are organized logically, but some part of the sample may not be fully developed  
     * Some transition of ideas is evident  
     * Meaning is conveyed but breaks down at times  
     * Mechanical errors are present but do not disrupt communication  
     * Shows a good understanding of writing and topic development |
| 3    | * Vocabulary is simple  
     * Organization may be extremely simple or there may be evidence of disorganization  
     * There are a few transitional markers or repetitive transitional markers  
     * Meaning is frequently not clear  
     * Mechanical errors affect communication  
     * Shows some understanding of writing and topic development |
| 2    | * Vocabulary is limited and repetitious  
     * Sample is comprised of only a few disjointed sentences  
     * No transitional markers  
     * Meaning is unclear  
     * Mechanical errors cause serious disruption in communication  
     * Shows little evidence of discourse understanding |
| 1    | * Responds with a few isolated words  
     * No complete sentences are written  
     * No evidence of concepts of writing |
| 0    | * No response |


The 6 point scale above includes broad descriptors of what a student’s essay reflects for each band. It is quite apparent that graders using this scale are expected to pay attention to vocabulary, meaning, organisation, topic development and communication. Mechanics such as punctuation are secondary to communication. A second example of a scoring guide, taken from the Test of Written English, is given in Appendix A.

Bailey (1998) also describes another type of scoring related to the holistic approach which she refers to as primary trait scoring. In primary trait scoring, a particular functional focus is selected which is based on the purpose of the writing and grading is based on how well the student is able to express that function. For example, if the function is to persuade, scoring would be on how well the author has been able to persuade the grader rather than how well organised the ideas were, or how grammatical the structures in the essay were. This technique to grading emphasises functional and communicative ability rather than discrete linguistic ability and accuracy.

A variation of this technique, multi-trait scoring (Hamp-Lyons, 1991), is also popular. In multi-trait scoring the scorer identifies several constructs that make up writing ability and uses bands and band scores to describe and score each construct. Although multi-trait scoring resembles analytical scoring discussed in the next section, analytical scoring usually assesses the more traditional elements of writing. In the example of multi-trait scoring provided in Appendix B, the focus is on specific aspects of how well a task is performed.
Analytical Scoring

Analytical scoring is a familiar approach to many teachers. In analytical scoring, raters assess students’ performance on a variety of categories which are hypothesised to make up the skill of writing. Content, for example, is often seen as an important aspect of writing – i.e. is there substance to what is written? Is the essay meaningful? Similarly, we may also want to consider the organisation of the essay. Does the writer begin the essay with an appropriate topic sentence? Are there good transitions between paragraphs? Other categories that we may want to also consider include vocabulary, language use and mechanics. The following are some possible components used in assessing writing ability using an analytical scoring approach and the suggested weightage assigned to each:

<table>
<thead>
<tr>
<th>Components</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content</td>
<td>30 points</td>
</tr>
<tr>
<td>Organisation</td>
<td>20 points</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>20 points</td>
</tr>
<tr>
<td>Language use</td>
<td>25 points</td>
</tr>
<tr>
<td>Mechanics</td>
<td>5 points</td>
</tr>
</tbody>
</table>

The points assigned to each component reflect the importance of each of the components.

Objective Scoring

A third type of scoring approach is the objective scoring approach. This scoring approach relies on quantified methods of evaluating students’ writing. A sample of how objective scoring is conducted is given by Bailey (1999) as follows:

Establish standardization by limiting the length of the assessment: Count the first 250 words of the essay

Identify the elements to be assessed: Go through the essay up to the 250th word underlining every mistake – from spelling and mechanics through verb tenses, morphology, vocabulary, etc. Include every error that a literate reader might note.

Operationalise the assessment: Assign a weight score to each error, from 3 to 1. A score of 3 is a severe distortion of readability or flow of ideas; 2 is a moderate distortion; and 1 is a minor error that does not affect readability in any significant way.

Quantify the assessment: Calculate the essay Correctness Score by using 250 words as the numerator of a fraction, and the sum of error scores as the denominator: The denominator is the sum of all the error scores:

\[
\frac{250 \text{ (words)}}{71 \text{ (sum of errors)}} = 3.52 \text{ correctness score}
\]

The steps described above help to provide a clear and systematic method for assessing essays. Objective scoring does not necessarily need to use the same values as in this example. The most important element in this approach is the objective scoring which is determined through the unbiased and fixed values provided according to some concrete aspect of the essay such as the number of mistakes made.

Comparing the three approaches. Each of the three scoring approaches claims to have its own advantages and disadvantages. These can be illustrated by Table 3 below.
Table 3. Comparison of the advantages and disadvantages of the three approaches to scoring essays

<table>
<thead>
<tr>
<th>Scoring Approach</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holistic</td>
<td>• quickly graded</td>
<td>• the single score may actually mask differences across individual compositions</td>
</tr>
<tr>
<td></td>
<td>• provide a public standard that is easily understood</td>
<td>• does not provide a lot of diagnostic feedback</td>
</tr>
<tr>
<td></td>
<td>• relatively higher degree of rater reliability</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• applicable to the assessment of many different topics</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• emphasise the students’ strengths rather than their weaknesses</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• grading a paper can be stopped and resumed with ease even after a period of time as the consistency or reliability of scoring is maintained</td>
<td></td>
</tr>
<tr>
<td>Analytical</td>
<td>• provides clear guidelines in grading in the form of the various components</td>
<td>• writing ability is unnaturally split up into components</td>
</tr>
<tr>
<td></td>
<td>• allows the graders to consciously address important aspects of writing</td>
<td></td>
</tr>
<tr>
<td>Objective</td>
<td>• objectivity in scoring</td>
<td>• still some degree of subjectivity involved</td>
</tr>
<tr>
<td></td>
<td>• grading a paper can be stopped and resumed with ease even after a period of time as the consistency or reliability of scoring is maintained</td>
<td>• accentuates negative aspects of the learner’s writing without giving credit for what they can do well</td>
</tr>
</tbody>
</table>

There are several obvious advantages of using the holistic or impressionistic method of scoring essays. Firstly, the test can be quickly graded as it only requires the grader to assess each essay holistically. Secondly, the scale can provide a public standard that is understood by the teachers and students alike. Another advantage is that because of the use of bands, there is bound to be a higher degree of rater reliability. Bailey (1998) also states that holistic scoring has the advantage of being applicable to the assessment of many different topics. Finally, because it is impressionistic, this grading method is thought to also emphasise the students’ strengths rather than their weaknesses.

There are also some clear disadvantages of the holistic method of scoring. Firstly, because there is only one score reported, the single score may actually mask differences across individual compositions. Students who are actually different may be included in the same band but at different ends of the band. The use of a single band score also does not provide a lot of diagnostic feedback if we were to compare against analytical scoring which can help students and teachers alike determine what aspect of the essay the student is weak at.

As far as the analytical approach is concerned, several advantages have been identified which are firstly that it provides clear guidelines in grading in the form of the various components and secondly that these components allow the graders to consciously address important aspects of writing. While
what the components of analytical scoring are have been suggested earlier, we should be aware that these components are intended to writing ability. Ideas about what writing is may change and this change can quite easily be included in analytical scoring by adding new components and deleting others.

Some educators believe that a major disadvantage of the analytical approach is that writing ability is unnaturally split up into components. Language is a whole, and by breaking it up into component parts, it no longer retains its true form. As such, assessment based on components is seen not to grade language as a whole but simply components of the language and nothing more. Grading based on the analytical approach is therefore seen as somewhat unrepresentative of actual writing ability.

The biggest advantage that the objective approach wishes to lay claim to is the objectivity in scoring. Because it uses a fixed method of scoring, proponents consider it to be unaffected by the biases of the graders. The objective approach is also seen to be very systematic and grading a paper can be stopped and resumed with ease even after a period of time as the consistency or reliability of scoring is maintained through the systematic approach.

The objective approach is not a very popular approach and the reasons why this is so is quite obvious. Firstly, despite its claims of being objective, there is still some degree of subjectivity involved. The weight scores assigned to each error, for example, is a subjective decision. Why three points deducted for a major error and not five points? A second and perhaps even more damaging criticism of this approach is that the error tally accentuates negative aspects of the learner’s writing without giving credit for what they can do well. It will therefore restrict any form of student creativity and experimentation – elements that are essential for the development of good writing ability.

**Ensuring accurate scoring of essays.** Because of the subjectivity involved in grading essays, many have given suggestions on how to ensure a greater degree of accuracy and consistency in scoring. The following diagram helps illustrate the various steps that need to be taken in order to ensure more accurate scoring of essays.

![Steps in ensuring accurate scoring of essays](image)

Familiarisation with a grading scale is an important step in achieving valid and accurate scores. If bands are used, there is an obvious need to fully understand what each band signifies. This stage of the grading process should therefore be given due consideration and not ignored. There are enough incidents of graders “jumping the gun” and assessing essays without first becoming familiar with the scoring criteria. This may only result in having to grade the paper again later on.
The purpose of identifying benchmark papers or anchor papers is to provide a clear and representative example of students’ work according to the grading criteria. Bands can only give a general description of what is expected. Anchor or benchmark papers provide concrete examples and help ensure fairness in grading.

When it comes to the actual grading, some recommend that we first quickly scan through all the essays and place them in stacks according to the bands on the scale. All papers which we consider A papers will be stacked together, the B papers will be placed together and so on. We can then read each paper more closely in order to confirm our initial impression. This procedure can be referred to as norming. If we need to assign more precise numerical scores, we can do so at this time. Another useful advice in grading essays, especially when students have been asked to write many essays, is to grade all the students on one essay first before moving on to the next essay. This is expected to help ensure more consistent grading.

Applications

Read the essay below. Score the essay according to the holistic approach and then according to the analytical approach (use the components discussed in this chapter). Were your two scores the same? If not, which approach produced the higher score? What kinds of problems did you encounter in using each scoring approach? Which approach do you prefer and why?

(Essay taken from the English of Malaysian School Students or EMAS corpus, Arshad Abd.Samad et al., 2002)
REFERENCES


Bateman, H. (2009). Some evidence to support the alignment of an LSP writing test to the CEFR. University of Cambridge ESOL Examinations Research Notes, 37, 29-34.


Ory, J. C. (n.d.) *Improving your test questions*. Office of instructional resources, University of Illinois at Urbana-Champaign.


