CHAPTER SIX

Scoring procedures for writing assessment

In the previous chapter, issues regarding the development and trialing of tasks for writing assessment were discussed. We now turn to the second key component of a writing assessment: procedures for scoring the written product. The scoring procedures are critical because the score is ultimately what will be used in making decisions and inferences about writers. As discussed in Chapter 4, a score in a writing assessment is the outcome of an interaction that involves not merely the test taker and the test, but the test taker, the prompt or task, the written text itself, the rater(s) and the rating scale (Hamp-Lyons, 1990; Kenyon, 1992; McNamara, 1996). Of these elements, two represent central considerations in scoring: defining the rating scale, and ensuring that raters use the scale appropriately and consistently. This chapter discusses these considerations and provides guidelines for designing rating scales, training raters, and ensuring the reliability and validity of scoring. For the purposes of this chapter, I will follow British and Australian usage and refer to the written text that is evaluated by raters as the 'script.' While this term is not widely used in the United States, I find it to be the easiest and most inclusive way to refer to the written response to a task on a writing test, whether it be an essay, a letter, or some other genre.

108
Table 6.1. Types of rating scales used for the assessment of writing

<table>
<thead>
<tr>
<th>Specific to a particular writing task</th>
<th>Generalizable to a class of writing tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single score</td>
<td>Primary Trait</td>
</tr>
<tr>
<td>Multiple scores</td>
<td></td>
</tr>
</tbody>
</table>

Rating scales

As McNamara (1996) notes, the scale that is used in assessing performance tasks such as writing tests represents, implicitly or explicitly, the theoretical basis upon which the test is founded; that is, it embodies the test (or scale) developer's notion of what skills or abilities are being measured by the test. For this reason the development of a scale (or set of scales) and the descriptors for each scale level are of critical importance for the validity of the assessment.

Types of rating scales

One of the first decisions to be made in determining a system for scoring is what type of rating scale will be used: that is, should a single score be given to each script, or will each script be scored on several different features? This issue has been the subject of a great deal of research and discussion over the past three decades. In the composition literature, three main types of rating scales are discussed: primary trait scales, holistic scales, and analytic scales. In recent second-language literature, a fourth type of scale, called a multiple-trait scale, is also frequently referred to (Hamp-Lyons, 1990; Cohen, 1994). However, many of the characteristics ascribed to multiple-trait scales have to do more with procedures for developing and using the scales, rather than with the description of the scales themselves. Thus for the purposes of this book, multiple-trait scales are not distinguished from analytic scales. The three types of scales can be characterized by two distinctive features: (1) whether the scale is intended to be specific to a single writing task or generalized to a class of tasks (broadly or narrowly defined), and (2) whether a single score or multiple scores are given to each script. Table 6.1 summarizes the three types of scales according to this scheme. As the
table shows, primary trait scales are specific to a particular writing task, while holistic and analytic scales can be used for grading multiple tasks. These three types of rating scales are discussed in more detail below.

**Primary trait scoring**

Primary trait scoring is most closely associated with the work of Lloyd-Jones (1977) for the National Assessment of Educational Progress (NAEP), a large-scale testing program for schools in the US. The philosophy behind primary trait scoring is that it is important to understand how well students can write within a narrowly defined range of discourse (e.g. persuasion or explanation). In primary trait scoring, the rating scale is defined with respect to the specific writing assignment and essays are judged according to the degree of success with which the writer has carried out the assignment. For each writing task in a primary trait assessment, a scoring rubric is created which includes: (a) the writing task; (b) a statement of the primary rhetorical trait (for example, persuasive essay, congratulatory letter) elicited by the task; (c) a hypothesis about the expected performance on the task; (d) a statement of the relationship between the task and the primary trait; (e) a rating scale which articulates levels of performance; (f) sample scripts at each level; and (g) explanations of why each script was scored as it was. A primary trait scoring guide can include several categories on which each script is to be judged. Part of a scoring guide for a primary trait assessment is found in Figure 6.1, taken from Lloyd-Jones (1977). As the figure shows, the scoring rubric is fairly detailed and very specific in terms of how different test takers approach the writing task. It is clear that primary trait scoring is very time- and labor-intensive, as a scoring guide must be developed for every writing task: Lloyd-Jones (1977) estimates that creating a scoring guide takes an average of 60 to 80 hours per task. For this reason alone, primary trait scoring has not been generally adopted in many assessment programs, even though it has the potential of providing rich information about students’ abilities, provided that enough samples of writing are collected from each student.

In second-language writing assessment, primary trait assessment has not been widely used, and little information exists on how primary trait scoring might be applied in second-language testing.
Directions: Look carefully at the picture. These kids are having fun jumping on the overturned boat. Imagine you are one of the children in the picture. Or if you wish, imagine that you are someone standing nearby watching the children. Tell what is going on as he or she would tell it. Write as if you were telling this to a good friend, in a way that expresses strong feelings. Help your friend FEEL the experience too. Space is provided on the next three pages.

**NAEP Scoring Guide: Children on Boat**

**Background**
*Primary Trait.* Imaginative Expression of Feeling through Inventive Elaboration of a *Point of View.*

**Final Scoring Guide**

**ENTIRE EXERCISE**
- 0 No response, sentence fragment
- 1 Scorables
- 2 Illegible or illiterate
- 3 Does not refer to the picture at all
- 9 I don’t know

**USE OF DIALOGUE**
- 0 Does not use dialogue in the story.
- 1 Direct quote from one person in the story. The one person may talk more than once. When in doubt whether two statements are made by the same person or different people, code 1. A direct quote of a thought also counts. Can be in hypothetical tense.
- 2 Direct quote from two or more persons in the story.

**POINT OF VIEW**
- 0 Point of view cannot be determined, or does not control point of view.
- 1 Point of view is consistently one of the five children. Include “If I were one of the children . . .” and recalling participation as one of the children.
- 2 Point of view is consistently one of an observer. When an observer joins the children in the play, the point of view is still “2” because the observer makes a sixth person playing. Include papers with minimal evidence even when difficult to tell which point of view is being taken.

**TENSE**
- 0 Cannot determine time, or does not control tense. (One wrong tense places the paper in this category, except drowned in the present.)
- 1 Present tense—past tense may also be present if not part of the “main line” of the story.
- 2 Past tense—If a past tense description is acceptable brought up to present, code as “past.” Sometimes the present is used to create a frame for past events. Code this as past, since the actual description is, in the past.
- 3 Hypothetical time—Papers written entirely in the “If I were on the boat” or “If I were there, I would.” These papers often include future references such as “when I get on the boat I will.” If part is hypothetical and rest past or present and tense is controlled, code present or past. If the introduction, up to two sentences, is only part in past or present then code hypothetical.

---

Figure 6.1 Primary trait scoring guide (Lloyd-Jones, 1977)
However, as Hamp-Lyons (1991a) points out, primary trait scoring might be particularly valuable for second-language learners in a school context, where parents, who are themselves not proficient in the language of the school, can benefit from a description of what their child is capable of doing with the language.

**Holistic scoring**

Many assessment programs rely on **holistic scoring**, or the assigning of a single score to a script based on the overall impression of the script. In a typical holistic scoring session, each script is read quickly and then judged against a rating scale, or scoring rubric, that outlines the scoring criteria. The existence of a scoring rubric distinguishes holistic scoring from its earlier, less reliable predecessor, **general impression marking**, in which criteria are never explicitly stated. The rubric is complemented by a set of anchor or benchmark scripts at each level that are intended to exemplify the criteria for that level, and raters are carefully trained to adhere to the rubric when scoring scripts. Note, however, that the existence of a rubric, benchmark scripts, and rater training is not limited to holistic scoring; on the contrary, these features are recognized as good practice in writing assessment, regardless of the type of scale used.

A well-known example of a holistic scoring rubric in ESL is the scale used for the TOEFL Writing Test, formerly known as the Test of Written English (TWE) (see Figure 6.2). As can be seen from the figure, the scale contains descriptors of the syntactic and rhetorical qualities of six levels of writing proficiency. Holistic scoring has become widely used in writing assessment over the past 25 years and has a number of positive features. From a practical standpoint, it is faster (and therefore less expensive) to read a script once and assign a single score than to read it several times, each time focusing on a different aspect of the writing. However, as White (1984, 1985), one of the leading proponents of holistic scoring, notes, there are also other advantages to holistic scoring. White maintains that holistic scoring is intended to focus the reader's attention on the strengths of the writing, not on its deficiencies, so that writers are rewarded for what they do well. Holistic scoring rubrics can be designed to focus readers' attention on certain aspects of writing, depending on what is deemed most essential in the context, and thus can provide important
6 An essay at this level
- effectively addresses the writing task
- is well organized and well developed
- uses clearly appropriate details to support a thesis or illustrate ideas
- displays consistent facility in use of language
- demonstrates syntactic variety and appropriate word choice though it may have occasional errors

5 An essay at this level
- may address some parts of the task more effectively than others
- is generally well organized and developed
- uses details to support a thesis or illustrate an idea
- displays facility in the use of language
- demonstrates some syntactic variety and range of vocabulary, though it will probably have occasional errors

4 An essay at this level
- addresses the writing topic adequately but may slight parts of the task
- is adequately organized and developed
- uses some details to support a thesis or illustrate an idea
- demonstrates adequate but possibly inconsistent facility with syntax and usage
- may contain some errors that occasionally obscure meaning

3 An essay at this level may reveal one or more of the following weaknesses:
- inadequate organization or development
- inappropriate or insufficient details to support or illustrate generalizations
- a noticeably inappropriate choice of words or word forms
- an accumulation of errors in sentence structure and/or usage

2 An essay at this level is seriously flawed by one or more of the following weaknesses:
- serious disorganization or underdevelopment
- little or no detail, or irrelevant specifics
- serious and frequent errors in sentence structure or usage
- serious problems with focus

1 An essay at this level
- may be incoherent
- may be undeveloped
- may contain severe and persistent writing errors

0 A paper is rated 0 if it contains no response, merely copies the topic, is off-topic, is written in a foreign language, or consists of only keystroke characters.

Figure 6.2 TOEFL writing scoring guide
information about those aspects in an efficient manner. White also argues that holistic scoring is more valid than analytic scoring methods because it reflects most closely the authentic, personal reaction of a reader to a text, and that, in analytic scoring methods, ‘too much attention to the parts is likely to obscure the meaning of the whole’ (White, 1984: 409).

On the other hand, holistic scoring has several disadvantages, particularly in second-language contexts. One drawback to holistic scoring is that a single score does not provide useful diagnostic information about a person’s writing ability, as a single score does not allow raters to distinguish between various aspects of writing such as control of syntax, depth of vocabulary, organization, and so on. This is especially problematic for second-language writers, since different aspects of writing ability develop at different rates for different writers: some writers have excellent writing skills in terms of content and organization but may have much lower grammatical control, while others may have an excellent grasp of sentence structure but may not know how to organize their writing in a logical way.

Another disadvantage of holistic scoring is that holistic scores are not always easy to interpret, as raters do not necessarily use the same criteria to arrive at the same scores: for example, a certain script might be given a 4 on a holistic scale by one rater because of its rhetorical features (content, organization, development), while another rater might give the same script a 4 because of its linguistic features (control of grammar and vocabulary). Holistic scores have also been shown to correlate with relatively superficial characteristics such as length and handwriting (Markham, 1976; Sloan and McGinnis, 1982). Holistic scoring has also come under criticism in recent years for its focus on achieving high inter-rater reliability at the expense of validity, as discussed in Chapter 3.

Analytic scoring

In analytic scoring, scripts are rated on several aspects of writing or criteria rather than given a single score. Depending on the purpose of the assessment, scripts might be rated on such features as content, organization, cohesion, register, vocabulary, grammar, or mechanics. Analytic scoring schemes thus provide more detailed information about a test taker’s performance in different aspects of writing and
are for this reason preferred over holistic schemes by many writing specialists.

One of the best known and most widely used analytic scales in ESL was created by Jacobs et al. (1981) (see Figure 6.3). In the Jacobs et al. scale, scripts are rated on five aspects of writing: content, organization, vocabulary, language use, and mechanics. The five aspects are differentially weighted to emphasize first content (30 points) and next language use (25 points), with organization and vocabulary weighted equally (20 points) and mechanics receiving very little emphasis (5 points). This scale has been adopted by numerous college-level writing programs, and is accompanied by training materials and sample compositions so that users can fairly quickly learn to apply the scale. A slightly different approach to analytic scoring for second-language writing assessment is a set of scales developed for the Test in English for Educational Purposes (TEEP) by Cyril Weir (1988), reproduced as Figure 6.4. Instead of a single scale composed of a number of subscales, Weir’s scheme consists of seven scales, each divided into four levels with score points ranging from 0 to 3. The first four scales are related to communicative effectiveness, while the others relate to accuracy. Like the Jacobs et al. scale, the TEEP scale was extensively piloted and revised to make sure that it could be applied reliably by trained raters. A third example of an analytic scoring system is the Michigan Writing Assessment Scoring Guide (Hamp-Lyons, 1990; 1991b) for grading an entry-level university writing examination (see Figure 6.5). The Michigan Writing Assessment is scored on three rating scales: Ideas and Arguments, Rhetorical Features, and Language Control. Like the TEEP scales, the three scales are not combined in a single score, but are reported separately and thus provide valuable diagnostic information to teachers and test takers. A distinguishing feature of this assessment is that the scales were locally developed in consultation with the university faculty, both within and outside of composition, and incorporate considerations of good writing as defined by a variety of constituents. As a result, Hamp-Lyons (1991b) states that the assessment has had a positive reception among students, faculty, advisors, and other community members because it reflects the concerns of and is easily interpreted by these varied constituencies. The examples of analytic scales presented here reflect an understanding that has become well established in writing assessment: that is, the importance of using an explicit and detailed scoring rubric. Criticisms of analytic scoring that
<table>
<thead>
<tr>
<th>SCORE</th>
<th>LEVEL</th>
<th>CRITERIA</th>
<th>COMMENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>30-27</td>
<td>EXCELLENT TO VERY GOOD: knowledgeable • substantive • thorough development of thesis • relevant to assigned topic.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24-22</td>
<td>GOOD TO AVERAGE: some knowledge of subject • adequate range • limited development of thesis • mostly relevant to topic, but lacks detail</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21-17</td>
<td>FAIR TO POOR: limited knowledge of subject • little substance • inadequate development of topic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16-13</td>
<td>VERY POOR: does not show knowledge of subject • non-substantive • not pertinent • OR not enough to evaluate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20-18</td>
<td>EXCELLENT TO VERY GOOD: fluent expression • ideas clearly stated/supported • succinct • well-organized • logical sequencing • cohesive</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17-14</td>
<td>GOOD TO AVERAGE: somewhat choppy • loosely organized but main ideas stand out • limited support • logical but incomplete sequencing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13-10</td>
<td>FAIR TO POOR: non-fluent • ideas confused or disconnected • lacks logical sequencing and development</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9-7</td>
<td>VERY POOR: does not communicate • no organization • OR not enough to evaluate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20-18</td>
<td>EXCELLENT TO VERY GOOD: sophisticated range • effective word/idiom choice and usage • word form mastery • appropriate register</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17-14</td>
<td>GOOD TO AVERAGE: adequate range • occasional errors of word/idiom form, choice, usage but meaning not obscured</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13-10</td>
<td>FAIR TO POOR: limited range • frequent errors of word/idiom form, choice, usage • meaning confused or obscured</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9-7</td>
<td>VERY POOR: essentially translation • little knowledge of English vocabulary, idioms, word form • OR not enough to evaluate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25-22</td>
<td>EXCELLENT TO VERY GOOD: effective complex constructions • few errors of agreement, tense, number, word order/function, articles, pronouns, prepositions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21-18</td>
<td>GOOD TO AVERAGE: effective but simple constructions • minor problems in complex constructions • several errors of agreement, tense, number, word order/function, articles, pronouns, prepositions but meaning seldom obscured</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17-11</td>
<td>FAIR TO POOR: major problems in simple/complex constructions • frequent errors of negation, agreement, tense, number, word order/function, articles, pronouns, prepositions and/or fragments, run-ons, deletions • meaning confused or obscured</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10-5</td>
<td>VERY POOR: virtually no mastery of sentence construction rules • dominated by errors • does not communicate • OR not enough to evaluate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>EXCELLENT TO VERY GOOD: demonstrates mastery of conventions • few errors of spelling, punctuation, capitalization, paragraphing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>GOOD TO AVERAGE: occasional errors of spelling, punctuation, capitalization, paragraphing but meaning not obscured</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>FAIR TO POOR: frequent errors of spelling, punctuation, capitalization, paragraphing • poor handwriting • meaning confused or obscured</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>VERY POOR: no mastery of conventions • dominated by errors of spelling, punctuation, capitalization, paragraphing • handwriting illegible • OR not enough to evaluate</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 6.3** Jacobs et al.'s (1981) scoring profile
A. Relevance and adequacy of content
   0. The answer bears almost no relation to the task set. Totally inadequate answer.
   1. Answer of limited relevance to the task set. Possibly major gaps in treatment of topic and/or pointless repetition.
   2. For the most part answers the tasks set, though there may be some gaps or redundant information.
   3. Relevant and adequate answer to the task set.

B. Compositional organisation
   0. No apparent organisation of content.
   1. Very little organisation of content. Underlying structure not sufficiently controlled.
   2. Some organisational skills in evidence, but not adequately controlled.
   3. Overall shape and internal pattern clear. Organisational skills adequately controlled.

C. Cohesion
   0. Cohesion almost totally absent. Writing so fragmentary that comprehension of the intended communication is virtually impossible.
   1. Unsatisfactory cohesion may cause difficulty in comprehension of most of the intended communication.
   2. For the most part satisfactory cohesion although occasional deficiencies may mean that certain parts of the communication are not always effective.
   3. Satisfactory use of cohesion resulting in effective communication.

D. Adequacy of vocabulary for purpose
   0. Vocabulary inadequate even for the most basic parts of the intended communication.
   1. Frequent inadequacies in vocabulary for the task. Perhaps frequent lexical inappropriacies and/or repetition.
   2. Some inadequacies in vocabulary for the task. Perhaps some lexical inappropriacies and/or circumlocution.
   3. Almost no inadequacies in vocabulary for the task. Only rare inappropriacies and/or circumlocution.

E. Grammar
   0. Almost all grammatical patterns inaccurate.
   1. Frequent grammatical inaccuracies.
   2. Some grammatical inaccuracies.
   3. Almost no grammatical inaccuracies.

F. Mechanical accuracy I (punctuation)
   0. Ignorance of conventions of punctuation.
   1. Low standard of accuracy in punctuation.
   2. Some inaccuracies in punctuation.
   3. Almost no inaccuracies in punctuation.

G. Mechanical accuracy II (spelling)
   0. Almost all spelling inaccurate.
   1. Low standard of accuracy in spelling.
   2. Some inaccuracies in spelling.
   3. Almost no inaccuracies in spelling.

Figure 6.4 TEEP attribute writing scales (Weir, 1990)
<table>
<thead>
<tr>
<th>Ideas and Arguments</th>
<th>Rhetorical Features</th>
<th>Language Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 The essay deals with the issues centrally and fully. The position is clear, and strongly and substantially argued. The complexity of the issues is treated seriously and the viewpoints of other people are taken into account very well.</td>
<td>The essay has rhetorical control at the highest level, showing unity and subtle management. Ideas are balanced with support and the whole essay shows strong control of organization appropriate to the content. Textual elements are well connected through logical or linguistic transitions and there is no repetition or redundancy.</td>
<td>The essay has excellent language control with elegance of diction and style. Grammatical structures and vocabulary are well-chosen to express the ideas and to carry out the intentions.</td>
</tr>
<tr>
<td>5 The essay deals with the issues well. The position is clear and substantial arguments are presented. The complexity of the issues or other viewpoints on them have been taken into account.</td>
<td>The essay shows strong rhetorical control and is well managed. Ideas are generally balanced with support and the whole essay shows good control of organization appropriate to the content. Textual elements are generally well connected although there may be occasional lack of rhetorical fluency: redundancy, repetition, or a missing transition.</td>
<td>The essay has strong language control and reads smoothly. Grammatical structures and vocabulary are generally well-chosen to express the ideas and to carry out the intentions.</td>
</tr>
<tr>
<td>4 The essay talks about the issues but could be better focused or developed. The position is thoughtful but could be clearer or the arguments could have more substance. Repetition or inconsistency may occur occasionally. The writer has clearly tried to take the complexity of the issues or viewpoints on them into account.</td>
<td>The essay shows acceptable rhetorical control and is generally managed fairly well. Much of the time ideas are balanced with support, and the organization is appropriate to the content. There is evidence of planning and the parts of the essay are usually adequately connected, although there are some instances of lack of rhetorical fluency.</td>
<td>The essay has good language control although it lacks fluidity. The grammatical structures used and the vocabulary chosen are able to express the ideas and carry the meaning quite well; although readers notice occasional language errors.</td>
</tr>
<tr>
<td>Ideas and Arguments</td>
<td>Rhetorical Features</td>
<td>Language Control</td>
</tr>
<tr>
<td>-----------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>3 The essay considers the issues but tends to rely on opinions or claims without</td>
<td>The essay has uncertain rhetorical control and is generally not very well managed.</td>
<td>The essay has language control which is acceptable but limited. Although the</td>
</tr>
<tr>
<td>the substance of evidence. The essay may be repetitive or inconsistent; the position needs to be clearer or the arguments need to be more convincing. If there is an attempt to account for the complexity of the issues or other viewpoints this is not fully controlled and only partly successful.</td>
<td>The organization may be adequate to the content, but ideas are not always balanced with support. Failures of rhetorical fluency are noticeable although there seems to have been an attempt at planning and some transitions are successful.</td>
<td>grammatical structures used and the vocabulary chosen express the ideas and carry the meaning adequately, readers are aware of language errors or limited choice of language forms.</td>
</tr>
<tr>
<td>2 The essay talks generally about the topic but does not come to grips with ideas about it, raising superficial arguments or moving from one point to another without developing any fully. Other viewpoints are not given any serious attention.</td>
<td>The essay lacks rhetorical control most of the time, and the overall shape of the essay is hard to recognize. Ideas are generally not balanced with evidence, and the lack of an organizing principle is a problem. Transitions across and within sentences are attempted with only occasional success.</td>
<td>The essay has rather weak language control. Although the grammatical structures used and vocabulary chosen express the ideas and carry the meaning most of the time, readers are troubled by language errors or limited choice of language forms.</td>
</tr>
<tr>
<td>1 The essay does not develop or support an argument about the topic, although it may ‘talk about’ the topic.</td>
<td>The essay demonstrates little rhetorical control. There is little evidence of planning or organization, and the parts of the essay are poorly connected.</td>
<td>The essay demonstrates little language control. Language errors and restricted choice of language forms are so noticeable that readers are seriously distracted by them.</td>
</tr>
</tbody>
</table>

Figure 6.5 Michigan writing assessment scoring guide

appeared twenty or thirty years ago pointed out quite rightly that scoring scripts on such features as ‘diction’ or ‘flavor’ was highly subjective because of the use of vague, indefinable criteria. Current scholarship (e.g. Weir, 1990; Hamp-Lyons, 1990; Alderson, 1991; Bachman and Palmer, 1996) emphasizes the need for clearly defined criteria and well-articulated levels for each scale or subscale within an analytic scoring scheme.

As mentioned above, the primary advantage of an analytic scoring
scheme over a holistic scheme is that it provides more useful diagnostic information about students' writing abilities. However, analytic scoring has a number of other advantages over holistic scoring as well. First, some research suggests that analytic scoring is more useful in rater training, as inexperienced raters can more easily understand and apply the criteria in separate scales than in holistic scales (Francis, 1977, and Adams, 1981, both cited in Weir, 1990). Analytic scoring is particularly useful for second-language learners, who are more likely to show a marked or uneven profile across different aspects of writing: for example, a script may be quite well developed but have numerous grammatical errors, or a script may demonstrate an admirable control of syntax but have little or no content (see Hamp-Lyons, 1991b, for a fuller discussion of this issue). Finally, analytic scoring can be more reliable than holistic scoring; just as reliability tends to increase when additional items are added to a discrete-point test, so a scoring scheme in which multiple scores are given to each script tends to improve reliability (Hamp-Lyons, 1991b; Huot, 1996).

The major disadvantage of analytic scoring is that it takes longer than holistic scoring, since readers are required to make more than one decision for every script. An additional problem with some analytic scoring schemes is that, if scores on the different scales are combined to make a composite score, a good deal of the information provided by the analytic scale is lost. It may also be the case that raters who are experienced at using a particular analytic scoring system may actually rate more holistically than analytically if scores are combined into a single score: experienced raters may target their ratings towards what they expect the total score to come out to be, and revise their analytic scores accordingly (Charlene Polio, personal communication, 1998).

To summarize, the choice about the kind of rating scale to use is not always clear-cut. A useful approach to making a decision is to appeal to the Bachman and Palmer (1996) framework of test usefulness discussed in Chapter 3. Table 6.2 presents a comparison of holistic and analytic scales based on the six qualities of test usefulness: reliability, construct validity, practicality, impact, authenticity, and interactivity. As Bachman and Palmer note, the choice of testing procedures involves finding the best possible combination of these qualities and deciding which qualities are most relevant in a given situation. For example, if large numbers of students need to be
Table 6.2. A comparison of holistic and analytic scales on six qualities of test usefulness

<table>
<thead>
<tr>
<th>Quality</th>
<th>Holistic Scale</th>
<th>Analytic Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliability</td>
<td>lower than analytic but still acceptable</td>
<td>higher than holistic</td>
</tr>
<tr>
<td>Construct Validity</td>
<td>holistic scale assumes that all relevant aspects of writing ability develop at the same rate and can thus be captured in a single score; holistic scores correlate with superficial aspects such as length and handwriting</td>
<td>analytic scales more appropriate for L2 writers as different aspects of writing ability develop at different rates</td>
</tr>
<tr>
<td>Practicality</td>
<td>relatively fast and easy</td>
<td>time-consuming; expensive</td>
</tr>
<tr>
<td>Impact</td>
<td>single score may mask an uneven writing profile and may be misleading for placement</td>
<td>more scales provide useful diagnostic information for placement and/or instruction; more useful for rater training</td>
</tr>
<tr>
<td>Authenticity</td>
<td>White (1995) argues that reading holistically is a more natural process than reading analytically</td>
<td>raters may read holistically and adjust analytic scores to match holistic impression</td>
</tr>
<tr>
<td>Interactivens*</td>
<td>n/a</td>
<td>n/a</td>
</tr>
</tbody>
</table>

*Interactive ness, as defined by Bachman and Palmer, relates to the interaction between the test taker and the test. It may be that this interaction is influenced by the rating scale if the test taker knows how his/her writing will be evaluated; this is an empirical question.

placed into writing courses in a limited time with limited resources, a holistic scale may be the most appropriate based on considerations of practicality. In this case, issues of reliability, validity, and impact can be ameliorated by the possibility of adjusting placements within the first week of class. On the other hand, a test of writing used for research purposes may have reliability and construct validity as central concerns, and practicality and impact may be of lesser significance. These issues must be resolved by the test users in considering all aspects of the situation.
Designing the scoring rubric

Factors to consider in designing a scoring rubric

Once it is determined what kind of rating scale is to be used, the next step is to design the scale, or scoring rubric, itself. The scoring rubric is critical, as it represents as explicitly as possible the definition of the skill(s) that the test is intended to measure, as mentioned above. However, it is not enough for a rubric to be clear and explicit: it must also be useable and interpretable, certainly by raters, and preferably by any and all stakeholders in the testing process, particularly test takers and decision makers. Factors to consider when designing a scoring rubric are as follows:

(a) **Who is going to use the scoring rubric?** Alderson (1991) notes that rating scales can have three distinct functions, depending on who is using them. **Constructor-oriented scales** are intended to guide the construction of tests at appropriate levels and thus include reference to the kinds of writing tasks that examinees would be expected to encounter, as in the following example from the American Council for the Teaching of Foreign Languages (ACTFL) writing scale: ‘Can write a personal letter on simple everyday topics or a simple report on an everyday event’ (ACTFL, 1985, cited in Alderson, 1991). **Assessor-oriented scales** are intended to guide the rating process, and focus on comparing the written text with descriptors on the scale. Finally, **user-oriented scales** are written with a focus on providing useful information to help test users interpret test scores. Alderson argues that it is important to be clear about the main function of a scale, and that problems are likely to arise when a scale intended for one purpose is used for another.

(b) **What aspect(s) of writing are most important, and how will they be divided up?** The Michigan Writing Assessment scale discussed above has a single category for language use. In some situations, however, in other situations it may be more appropriate to have separate scales for vocabulary use and grammatical accuracy. More detailed information about various aspects of language use would be particularly appropriate when the focus of the assessment is on the acquisition of specific language subskills, such as in low-proficiency non-academic classes or general foreign-language instruction. In
post-secondary academic settings, and specifically when a test is being used to certify proficiency for academic study or to exempt students from composition courses, more emphasis will be placed on communicative effectiveness and less on specific language features, making a general category for language use more appropriate.

(c) How many points, or scoring levels, will be used? It is important to be able to distinguish between writers of different abilities; however, there are limits to the number of distinctions that raters can reliably make. Many large-scale assessment programs such as the TOEFL use a six-point scale; however, others have used nine-point scales with success. Part of this decision is determined by the range of performances that can be reasonably expected of the population of test takers; another consideration is the use to which the test will be put. If the test is being used primarily to make pass/fail decisions (as in a university writing competence examination, for example) fewer score points may be needed. If the test will be used to place students into different courses, on the other hand, more score points will be needed. Bachman and Palmer (1996) recommend using more score points than there are decisions to be made, since ratings are never completely reliable; that is, since independent raters will not necessarily agree on exact scores. The background and experience of raters may also influence the number of score points to be used: more experienced raters may be able to make finer distinctions between scripts than less experienced raters, and thus may be able to make use of more score points on a scale reliably.

Some questions about scale points can only be determined through empirical means in pre-testing, by trying out the scale with a wide variety of scripts and raters to determine whether raters are able to use the entire range of scores and distinguish between scale steps reliably. For example, if pre-testing showed that raters were only using four out of six scale points on a scale, scale developers might consider eliminating the other two scale points altogether or rewording the descriptors so that distinctions between the levels would be easier for raters to make. Pollitt (1990) points out that the number of points on a rating scale that can be distinguished reliably is a function of the overall reliability of the test, claiming that it is overly optimistic to expect a test of writing to be able to distinguish reliably between five scale points or more.