(d) How will scores be reported? If an analytic score is used, scale scores can either be combined for a total score, as is done in the Jacobs et al. (1981) scale, or scale scores can be reported separately for diagnostic purposes. Combining scores has the obvious advantage of providing a single score to be used in decision-making (for example, for cut-off scores for placement, exit, or exemption purposes), and combined scores tend to be more reliable than the separate part scores. However, combining scores also has the effect of negating the principal advantage of an analytic scale, which is to provide richer information about students' abilities. Since the same score can be achieved by people with varying profiles (for example, strong in syntax, weak in rhetoric, or vice versa) a composite score is difficult to interpret, except for those test takers who earn very high or very low scores.

Reporting separate scores provides more useful diagnostic information and generally provides a more accurate picture of test takers' abilities in writing. However, separate scores can be harder for test users to interpret quickly and cannot be combined easily with other parts of a test battery for decision-making purposes. For this issue the audience for test results must be kept in mind: program administrators who have to make quick decisions about many students may prefer a single score, while advisors helping students decide which of several courses to take would probably prefer the richer information provided by separate scores.

A related issue is whether, and how much, to weight different components of writing ability in an analytic scale. Jacobs et al. (1981) have differential weightings of the various components already built into their scale, with content receiving the most weight and mechanics the least. Hamp-Lyons (1991b), on the other hand, recommends weighting all components equally, suggesting that if one component is to be weighted more heavily than others in a given context, a focused holistic scale might be more appropriate. Weighting of scores has two complementary but distinct aspects that must be taken into consideration: it represents an explicit statement of a theory of writing ability (i.e. that certain aspects are more or less important/relevant/involved than others), and it also has consequences for the final scores that are the basis for decisions. The effects of weighting on final scores depend not just on the weights given to each component, but also on other statistical factors, such as the amount of variation within each component, and the correlations, or relationships between com-
ponents. For this reason, if weighting is to be considered, the advice of a statistician familiar with these issues should be sought.

**Writing scale descriptors**

Once these issues have been resolved, at least preliminarily, the descriptors for the various levels of the scale itself can be written. This can be done *a priori*, by defining in advance the ability being measured and then describing a number of levels of attainment, from none to complete mastery. This is the approach advocated by Bachman and Palmer (1996) and shown in Figure 6.6, from a set of scales used for placement into a university writing program. The advantage of this approach, according to Bachman and Palmer, is that it allows one to make inferences about a test taker’s language ability on an absolute scale rather than relative to other test takers or to native speakers. However, a potential problem with this approach is that the scale descriptors tend to make imprecise distinctions between the levels (e.g. ‘excellent,’ ‘very good,’ ‘good,’ and so on). It is likely that inexperienced raters may have difficulties making these distinctions reliably without extensive training and repeated exposure to texts that instantiate the various scale levels. Another approach is to generate scale descriptions empirically, through the examination of actual scripts and/or operational ratings of writing performances. North and Schneider (1998) describe five data-based methods of scale construction that involve expert judgements of the key features at different levels of performance, statistical analysis of ratings vis-à-vis scale descriptors, or textual features of performances at different levels. The most common of these methods involves gathering sample scripts on a prototype writing task from students at all relevant levels of proficiency and, with a group of instructors familiar with the proficiency levels, defining the characteristics that differentiate the samples. In this case the definitions ordinarily take the form of verbal description rather than levels of mastery, although there is certainly an explicit rank ordering of these descriptions in terms of quality. These descriptions frequently refer to such notions as audience awareness, overall communicative effectiveness, and effect on the reader – notions which do not lend themselves easily to the levels-of-mastery approach. The rating scales in Figures 6.2 and 6.4 above were developed in this manner. North and Schneider (1998) caution that this method of scale
<table>
<thead>
<tr>
<th>Levels of ability/mastery</th>
<th>Description</th>
</tr>
</thead>
</table>
| 0 None                   | *No evidence of knowledge* of syntax  
  Range: zero  
  Accuracy: not relevant |
| 1 Limited                | *Limited knowledge* of syntax  
  Range: small  
  Accuracy: poor, moderate, or good accuracy. If test taker only attempts a very few structures, accuracy may be good. |
| 2 Moderate               | *Moderate knowledge* of syntax  
  Range: medium  
  Accuracy: moderate to good accuracy within range. If test taker attempts structures outside of the range that is controlled, accuracy may be poor. |
| 3 Extensive              | *Extensive knowledge* of syntax.  
  Range: large, few limitations  
  Accuracy: good accuracy, few errors |
| 4 Complete               | *Evidence of complete knowledge* of syntax  
  Range: evidence of unlimited range  
  Accuracy: evidence of complete control |

Figure 6.6 Scale: knowledge of syntax (Bachman and Palmer, 1996)

construction, which relies on expert judgement rather than statistical analyses, may be most useful for limited, specific contexts rather than for very large-scale (i.e. statewide or international) assessments.

The choice between *a priori* and empirical development of rating scales may depend in part on philosophical orientation – in particular, the degree to which one believes that the most important aspects of the ability being tested can be measured on a scale of none to complete mastery – and also on factors related to the purpose of the assessment. The mastery approach advocated by Bachman and Palmer (1996) is particularly appropriate if the construct is conceived of as an inherent ability (i.e. a student ‘has’ ability X) rather than in terms of a pragmatic ascription (i.e. a student ‘can do’ X, without reference to the exact nature of the underlying ability). The mastery approach is useful when we want to make inferences about underlying abilities, as in diagnostic
tests, while the pragmatic approach is useful when we want to predict how someone will perform on future similar tasks. One disadvantage of the pragmatic approach, according to Bachman and Palmer, is its inadequacy for making generalizations: if we are simply predicting how one will perform in the future on the basis of specific performance without referring to an underlying ability, we may not be able to generalize that performance to other task types.

However, the different scale levels are described, these characteristics form the basis for a preliminary scoring rubric against which further scripts can be judged. At this point, descriptors can be added, deleted, or modified until raters can agree on scores in the majority of cases. If the scoring rubric is to be used with a number of different writing tasks, the next step would be to apply the rubric to a variety of tasks to make sure that the descriptors are relevant and applicable to these new tasks. It is also helpful to devise a training procedure at this point to make sure that new raters are able to learn to apply the scoring rubric appropriately. Training procedures are discussed at length in the next section of this chapter.

It should be noted that the process of developing a scoring rubric, as in all aspects of test development, is iterative: it may well be that decisions about the aspects of writing to be scored and the number of score points for each aspect will need to be revisited one or more times as a result of trying out the rubric on actual scripts. For example, it may be that raters can reliably distinguish between five levels for one aspect of writing, but only four levels for another, or it may be that raters cannot reliably distinguish between two aspects, which may need to be collapsed into a single category.

Calculating total scores

Before the scoring rubric can be finalized, decisions need to be made about calculating reported scores, in particular, if and how a total score will be derived from individual raters’ scores, how much tolerance for discrepancies in ratings will be allowed, and what procedures will be followed in adjudicating discrepancies. For the purposes of this discussion, it is assumed that two raters will read and score each script independently of each other, with a third, senior rater, reading the sample in case of discrepancies.

In the simplest case, when the two raters are in agreement, the
reported score can be the sum or the average of the two raters’ scores. The practice of combining scores in these ways has the effect of doubling the possible range of scores (for example, if a six-point scale is used, the reported scores will range from 2 to 12 if scores are added together, or from 1 to 6, maintaining the original scale, with half-point intervals, if the scores are averaged).

Another factor that must be decided is the amount of variability among ratings that will be considered acceptable. Common practice on a six-point scale is that ratings more than one score point apart (e.g. one score of 3 and one of 5) are considered discrepant and must be resolved. In some testing programs where there is a specific pass/fail boundary (e.g. a score of 4 on a 6-point scale is passing, while a score of 3 is failing), if the two scores fall on either side of the boundary a third rater must resolve the discrepancy, even if the scores are only a point apart. Procedures for incorporating a third rating in the reported score vary from institution to institution: in some cases, the reported score is the average of the two closest scores, while in others all three scores are averaged. In the case of a third rater adjudicating at a pass/fail boundary, the third rater’s judgement will normally prevail and the two scores on the same side of the boundary will be used in calculating the final score.

**The scoring process**

Once the scoring rubric has been finalized, the next step is to select raters and design a process for the operational scoring of scripts. Much of the literature on large-scale testing has been concerned with procedures for obtaining satisfactory levels of reliability in an efficient manner; these procedures are introduced in this section.

**Procedures for assuring reliability**

In Chapter 3, several possible sources of unreliability were discussed, and in Chapter 5 I discussed procedures for assuring consistency or reliability of tasks or prompts. In this section I will discuss another potential source of unreliability in writing assessment: inconsistencies in scoring. Of these, there are two main types: (1) inconsistencies in the ratings of a single scorer across different scripts of similar
quality or the same script on different occasions, and (2) inconsist-
ences between different scorers. Both of these potential sources of
unreliability can be addressed by the procedures described in this
section. Procedures for estimating the effects of these two kinds of
scorer inconsistency on test scores are discussed later in this
chapter.

White (1984) outlined six practices and procedures that are impor-
tant for maintaining high reliability in large-scale assessments. Two
of these have already been discussed: the use of a scoring rubric that
details explicitly the criteria to be used in scoring, and the use of
sample scripts in training that exemplify points on the scale. It is only
when all raters are in close agreement on the scores for these sample
scripts that reliable scoring can take place. White’s other recomme-
dnations are as follows:

• Each script must be scored independently by at least two raters,
with a third rater adjudicating in cases of discrepancy.

• Scoring should be done in a controlled reading, by which is meant
that a group of readers meets together to grade scripts at the same
place and time. Two advantages of controlled reading are that the
circumstances under which scripts are read are controlled, thus
eliminating unnecessary sources of error variance, and that a posi-
tive social environment is formed which helps to enforce and main-
tain the rating standards. Unfortunately, group scoring is not
always feasible; Alderson et al. (1995: 133–5) present alternatives
for scoring when this is the case.

• Checks on the reading in progress by reading leaders (sometimes
called Table Leaders) help to ensure that individual readers are
maintaining the agreed-upon standards for grading.

• Evaluation and record keeping are essential for an ongoing assess-
ment program so that reliable readers are kept on and unreliable
readers are retrained or dropped if necessary.

White also makes the important point that the tone set by the reading
leaders has a tremendous influence on the success of the reading. If a
reading is led with sensitivity and respect, it can be an enjoyable and
professionally valuable experience for readers; on the other hand,
poorly run readings, in which readers feel exploited or coerced, can
turn readers against the grading process, which in turn can have
negative effects on the scoring itself.
Rater training

Building on White's recommendations and procedures followed in large-scale writing assessment sessions such as those conducted by Educational Testing Service, the following process is recommended for instituting rater training for writing assessment. Specific circumstances will dictate to what extent this exact process can be followed, but the major elements should be part of any rater training. For example, if a small number of scripts are to be read and only two or three raters are involved, a less complex procedure might be appropriate. Similarly, if the same raters are used for frequent rating sessions, a full-scale rater training session might not be needed for every test administration.

The leader (or preferably a team) should read through the scripts to find anchor/benchmark scripts that exemplify the different points on the scale. Depending on the size of the reading, the complexity of the scale, the number of readers, and their experience, anything from three to ten sets of anchor scripts may be required. It is also helpful to include in the training sets scripts that exemplify certain problematic situations, for example, scripts that do not respond to the task or simply copy the prompt, or scripts that represent the borderline between two critical levels (e.g. passing and failing).

The first set of scripts is generally given to raters in order (from highest to lowest or vice versa) with the appropriate scores indicated, and should be as unambiguous a set as possible. This set is used to familiarize readers with the scale and to instantiate certain features of the rubric. The leader can use these scripts to describe for the readers what is meant by phrases in the rubric (e.g. 'appropriate introduction,' 'minor errors'). At this and every point questions are raised by readers and discussed with the whole group.

Once the readers feel comfortable with the scale as defined by the leader and instantiated in the first set or sets of anchor scripts, a set can be given that includes one script at each level in random order. Readers can be told that there should be one script at each level and given a chance to rate the scripts themselves. Once readers are able to handle this task, more problematic sets can be given out, which may have more than one script at a given level or may be less clearly representative of certain points of the scale.

It should be noted that it is virtually impossible to get a large group of raters to agree on exact scores and that some disagreement among
raters is inevitable. In training raters, it is important to communicate to raters the amount of variability that is acceptable and to let them know that they are not required to be perfectly accurate at all times. Raters who consistently rate higher or lower than the rest of the group should be given feedback and perhaps additional training to bring their scores into alignment with the rest of the group.

In a large reading, additional retraining may be required at certain points. For example, if the reading takes place over more than one day, one or two sets of anchor scripts can be used to recalibrate readers each day. Similarly, if raters will be reading more than one topic, anchor scripts for each topic should be used if possible.

Once live rating is under way, it is important to ensure that scoring is independent— that is, that raters do not see and therefore cannot be influenced by scores given by other raters—and, of course, procedures for maintaining independence of scores must be devised well in advance. These procedures can be as complex as assigning different codes for scores to each rater or using invisible ink, or as simple as designing a rating sheet that can be folded in half once the first score is given so that it is hidden from the second rater. Whatever procedure is used, it is essential for the integrity of the scoring process that raters arrive at their scores independently, without reference to scores given by other raters. For this reason it is also important that raters do not write comments or underline errors when scoring scripts, to avoid influencing the scores given by other raters.

**Special problems in scoring**

In this section, some common problems in scoring writing are discussed. In an ideal world, writing prompts would be written so that every test taker could understand exactly what was required and would respond appropriately within his or her ability level. In the same ideal world, every test taker would agree to go along with the task set by the test writers in exactly the way the test writers envision. However, in the real world these conditions are virtually impossible to meet. It then becomes incumbent upon the test writers (or readers, if they are different people) to determine to what extent fulfilling the exact task set forth in the writing prompt should be part of the scoring procedure. While it is impossible to foresee every problem that might occur, it is advisable to anticipate as far as possible the kinds of
problems that might occur with a given prompt, to reduce the possibility that different raters will approach problematic scripts differently and thus introduce unwanted errors into the scoring procedures. Three types of problematic scripts will be discussed in this section: (1) scripts that are complete but do not address the intended task or fail to address parts of the task, (2) scripts that have clearly been written from memory rather than in response to the prompt, and (3) incomplete scripts – that is, scripts in which the writer has demonstrated an understanding of the important features of the task but was unable to complete the task in the allotted time (for example, the conclusion may be missing).

**Off-task scripts**
A writing prompt ordinarily contains specific instructions to the test taker about what topic to address and how to approach the topic – for example, compare and contrast two things, outline the causes of a particular problem, or give advantages and disadvantages of some course of action. One problematic area in scoring is a script that misconstrues these instructions – listing solutions to a problem rather than the causes, for instance, or providing only advantages and not disadvantages. In such cases the raters need to decide the extent to which task fulfillment is essential to the scoring. Decisions about task fulfillment are based on the purpose of the assessment and the type of scoring that is used. For the TOEFL, for example, the test writers and users are interested primarily in a general sense of a person's ability to create a coherent written text, not the quality of the ideas or the persuasiveness of the essay. In addition, the TOEFL Writing Test is intended to test writing and not reading ability. For these reasons, raters are told that an essay can be considered on-topic and ratable if it can at all be reasonably construed to have something to do with the prompt. For example, if the prompt required test takers to discuss the advantages of books over movies, an essay that simply discussed a recent book that the writer had read would not be considered off-topic. On the other hand, in primary trait scoring, task fulfillment is central to the scoring guide, and a script that did not address the demands of the task would receive a low score no matter how beautifully written it was. This issue harks back to the distinction between the strong and weak senses of performance testing discussed in Chapter 3.

From these examples, it is clear that what constitutes task fulfill-
ment lies in the purpose of the assessment and the type of scoring that is used. For example, with assessments whose main purpose is obtaining a ratable sample of writing that will demonstrate control of syntax and vocabulary, the degree to which writers follow the instructions exactly will be less important than for assessments whose purpose is to assess writers' ability to successfully communicate in writing, such as is typically of interest in academic settings.

**Memorized scripts**

Determining task fulfillment is particularly problematic in the case of memorized scripts: if it becomes generally known that readers are merely looking for a scorable sample of writing rather than the ability to accomplish a specific task through writing, some test takers may prepare for the test by memorizing a well-written script and writing it out from memory during the test. Obviously a memorized script does not provide an accurate sample of a test taker's ability, since there is no way of knowing the author or source of the script. Thus, steps must be taken to avoid the possibility of memorized scripts being proffered or accepted for scoring. This argues for requiring at least some adherence to the task in order for a script to be acceptable, although again it is a matter of degree, depending on the circumstances. Generally, the more specifically a writing test is geared to a particular situation, the more important the notion of task fulfillment becomes. For example, task fulfillment may be less important in a general proficiency test for placement into university-level foreign-language classes, where students may have no genuine need for using the language outside of the classroom, while it becomes much more critical in a vocational or professional LSP course that is preparing students to use English in their professions.

**Incomplete responses**

Another issue that must be resolved is that of incomplete responses. This issue becomes particularly pertinent in cases where task fulfillment is an integral part of the scoring rubric. If a test taker makes a promising beginning to a script but does not complete it, the reader's dilemma is whether to score the script based on the strength of what has been written, assuming that the writer could have made an appropriate conclusion given enough time, or whether to adhere strictly to the wording of the rubric and score the script on the basis of what is present in the script. Again, to resolve this dilemma one must
consider the purpose of the assessment and the impact that decisions
will have on test takers. In a fairly low-stakes test such as a placement
test, where additional information can be gathered after the fact to
support or revise a decision, one might decide in favor of the former
approach. On a higher-stakes test, on the other hand, especially in an
academic context where a writing test is supposed to be indicative of
other types of academic writing, one can be stricter about the rating
criteria and judge incomplete scripts more severely.

The fact that it is impossible to know ahead of time which of these
or other unanticipated problems will occur during a live reading
makes the pre-testing of prompts, as discussed in Chapter 5, critical.
Sample scripts on a topic are essential for providing information
about the variations in responses that can inform rater training, and it
is important for scoring leaders to be aware of and anticipate the
kinds of problems that may arise during the reading. For example,
scoring leaders should be prepared with advice on how to deal with
poor handwriting, extremely brief responses, or uncreative or sim-
plistic responses. When unanticipated problems arise during opera-
tional rating, it may be worthwhile keeping a written record of the
problem and its resolution so that raters are consistent in dealing
with the problem should it recur later on. Careful consideration of the
implications of decisions regarding problem scripts is critical to en-
suring a successful scoring session.

Evaluating scoring procedures

As with other aspects of testing, scoring procedures can be evaluated
according to the criteria of test usefulness outlined in Bachman and
Palmer (1996). In particular the aspects of reliability, validity, and
practicality need to be evaluated with reference to scoring proce-
dures. Of these qualities of usefulness, procedures for evaluating
reliability are the best known.

Assessing reliability of scores

There are a number of ways to investigate the reliability or consis-
tency of raters. Two important aspects of reliability are intra-rater
reliability (self-consistency) and inter-rater reliability (agreement
between raters). Intra-rater reliability refers to the tendency of a rater to give the same score to the same script on different occasions, while inter-rater reliability refers to the tendency of different raters to give the same scores to the same scripts. In the simplest cases, the reliability of ratings on a holistic scale between two raters, or between the scores given to the same samples by one rater on two different occasions, can be calculated by means of a correlation coefficient. This statistic is a number between 0 and 1 indicating the strength of the relationship between two sets of scores. A correlation coefficient close to 0 indicates that there is little or no relationship between the scores given by the first rater and those given by the second (or on the second occasion), while a coefficient close to 1 indicates a strong relationship between the sets of scores. The formulas for calculating the appropriate statistic, either the Spearman rank-order correlation coefficient or the Pearson product-moment correlation coefficient, can be found in any standard textbook on statistics, and reliability coefficients are easily calculated on readily available spreadsheet software such as Microsoft Excel.

A complementary approach to investigating inter-rater reliability, particularly when more than two raters are involved, is through the analysis of variance (ANOVA). ANOVA can be used to compare the distribution of scores given by a set of raters (assuming they have all scored the same scripts). The two main statistics used to describe the distribution of scores are the mean, or average score, and the standard deviation, or the average amount that scores differ from the mean. ANOVA can be used to determine whether there is any statistical difference between the mean scores of raters; that is, if some raters tend to give higher or lower scores than other raters, irrespective of the correlation among raters' scores. Further information about analysis of variance can be found in a textbook on statistics; useful discussions of both intra- and inter-rater reliability can be found in Alderson et al. (1995), Hatch and Lazaraton (1991), and Bachman (forthcoming).

Beyond inter-rater reliability, there are several ways of looking at rater agreement. To judge the overall success of a rating session, the scores given by the first and second readers can be cross-tabulated, as in Table 6.3. White (1984) states that an average reading using a six-point scale will have 7–10% of the ratings more than one point apart, while in an excellent reading only 5% of ratings will be discrepant. In the table, numbers outside the shaded area represent discrepant
scores, or cases where the two readers were more than one point apart. A total of 11 scores, or slightly over 9% of the total number of scores, are discrepant in this example, indicating an acceptable degree of rater disagreement, according to this, admittedly rule-of-thumb, standard. For an overall indication of rater agreement using a table such as Table 6.3, a coefficient of relationship (called kappa) can be derived, and this can be interpreted much the same way as a correlation coefficient, described above. The reader is referred to the references listed above for further details about this statistic. For individual readers, statistics can be compiled to include the number of scripts read, the number of scripts read that needed to be adjudicated by a third reader, and the number of times discrepancies were resolved in the reader’s favor. This information can be useful for providing feedback to readers regarding their accuracy and efficiency or for determining whether to reuse raters in future readings.

Assessing validity of scoring procedures

In addition to considering the reliability of scores, it is important to investigate how scoring procedures affect the construct validity of a writing test – that is, the validity of inferences made on the basis of test results. Looking at construct validity in terms of scoring procedures is somewhat less straightforward than investigating reliability, however, as it involves investigation of a multitude of factors from a variety of perspectives. One can start by asking a few basic questions.
First, do the scoring procedures – in particular, the scoring guide – accurately reflect the construct being measured? As discussed previously, the scoring guide represents an explicit statement of what aspects of writing are being considered as part of the construct, and thus the first question is whether that is in fact the case. For example, if we are interested primarily in accuracy of content and logical organization in writing, the scoring guide should not focus heavily on grammar.

Another important question is whether the scoring procedures are being implemented in an appropriate way. This question relates to rater behavior: if raters are not basing their judgements on the scale as defined in the scoring guide, questions must be asked about what is actually being measured, since the scoring guide, as mentioned above, represents the construct. The studies of rater behavior cited in Chapter 4, particularly those using think-aloud protocols to investigate raters’ decision-making processes, relate to this question and can serve as models for similar studies on a local level. Alternatively, a debriefing session following rating can be held to discuss how raters are interpreting the scale or any problems or concerns that may have arisen during the scoring session. Information from these sessions can then feed into further refinements of the scoring guide.

A third question is whether the scores obtained from the test allow us to make appropriate inferences about writing ability and thus appropriate decisions about test takers. To answer this question, we thus need to evaluate our scoring procedures in terms of their effect on outcomes. Outcomes can be assessed both in terms of consequences for individual test takers and in terms of consequences for the educational system or teaching program (cf. the discussion of washback in Chapter 3). For individuals, one way to assess the consequences of scoring procedures is to investigate the appropriateness of decisions made on the basis of the test scores. For example, if a writing test is being used for placement, the percentage of placements that were later changed based on additional information can be calculated, thus giving an indication of how well the test is functioning for this purpose.

On a systemic level, scoring procedures can have an effect (positive or negative) on instruction in a number of ways. First, disseminating the rating criteria among teachers, students, and other stakeholders allows for frank discussion of, and ideally consensus about, the goals
of writing instruction and the expected outcomes for students. Dissemination of rating criteria can lead to instructional changes, if teachers gear their instruction towards the aspects of writing that are seen as valued because they are emphasized in the criteria. By the same token, students may be motivated to work on those aspects of writing if they are aware of how their writing will be scored.

The type of rating scale used – holistic, analytic, or primary trait – is also relevant in this regard. Using an analytic scoring guide rather than a holistic one can have beneficial effects as it provides more information about the strengths and weaknesses of students and may allow instructors and curriculum developers to tailor instruction more closely to the needs of their students.

Evaluating the practicality of scoring procedures

As discussed in Chapter 3, the practicality of a test is a matter of the availability of resources vis-à-vis the resources required to develop, administer, and score the test. Given a finite amount of resources for testing, it may be necessary to optimize the allocation of resources to reach the desired levels of the various aspects of test usefulness. This may take the form of a cost–benefit analysis. For instance, one may believe that an analytic scale may provide more useful information about test takers’ abilities than a holistic scale, but the time required to provide analytic scores may not be worth the added benefits. As another example, if inter-rater reliability is not sufficiently high, several options might be explored: increase the number of writing tasks, increase the number of raters, or devote more time to refining the rating scale and training raters. The costs and benefits for each option will have to be weighed against the availability of resources.

On a more mundane level, following an operational scoring session, it is generally a good idea for scoring leaders to meet with raters to discuss ways in which the scoring process could be streamlined without sacrificing reliability or validity: could training have been more efficient, for example, or was the system for tracking scripts and reporting scores too cumbersome? Even when a scoring system has been in place for a long period of time, it is useful to hold such debriefing sessions regularly, as new ideas and new technologies (see Chapter 10) may arise that can be useful for increasing the efficiency of scoring.
Summary

This chapter has provided a discussion of scoring procedures for writing assessment, focusing on the use and development of rating scales and procedures for training, monitoring, and evaluating raters. The next chapter brings together many of the issues raised in this chapter and the previous one by presenting both the writing tasks and scoring procedures for several illustrative writing tests for different purposes.

Many of the procedures described in this chapter were developed originally to counter the criticism that subjective scoring of writing could never be reliable and that it thus was, if not impossible, at least highly impractical, to test writing by having test takers actually write, particularly in large-scale assessment. A great deal of attention has therefore, and with good reason, been focused on increasing reliability of scoring. Recently, however, the focus has shifted towards ensuring the validity of scoring, an issue that will be revisited in Chapter 10.

It is also becoming increasingly clear that, while timed impromptu writing has its place in writing assessment, for many purposes a single script written under time restrictions is insufficient, as it provides a very restricted picture of what test takers are able to accomplish with writing in the real world. Furthermore, timed writing tests, while useful for large-scale assessment, are of less value in classroom-based assessment of writing, for reasons that are discussed in Chapter 8. In Chapters 9 and 10, therefore, we will turn from timed writing tests to alternative forms of writing assessment: classroom-based writing evaluation, and portfolio assessment.