however, item analysis is performed simply to investigate how well the items on a test are working with a particular group of students. Item analysis can take numerous forms, but when testing for norm-referenced purposes, there are three types of analyses that are typically applied: item format analysis, item facility analysis, and item discrimination analysis. In developing CRTs, three other concerns become paramount: item quality analysis, the item difference index, and the *B*-index for each item.

GUIDELINES FOR ITEM FORMAT ANALYSIS

In analyzing **item format**, testers focus on the degree to which each item is properly written so that it measures all and only the desired content. Such analyses often involve making judgments about the adequacy of item formats. Consider the following multiple-choice grammar item:

The apple is located somewhere on or around _____.

(A) a table
(B) an table
(C) the table
(D) table

This item has two possible answers (A and C), is wordier than it needs to be ("located somewhere...or around" may be difficult, distracting, and superfluous), and repeats the word "table" inefficiently. Item format analysis could lead us to correct these problems and produce a better item as follows:

Do you see the chair and table? The apple is on _____ table.

(A) a
(B) an
(C) the
(D) (no article)

Now, the first sentence makes "the" the only correct answer; the item has been reworded to avoid difficult, distracting, and superfluous words; and the word *table* is moved up into the main part of the item so it is not repeated four times in the A-D options. The item may still be imperfect because other teachers have not given feedback on it, but it is considerably better than it was when first written.

The guidelines provided in this chapter are designed to help teachers make well-informed and relatively objective judgments about how well items are formatted. The first set of guidelines is a very general set that teachers can apply to virtually all types of items. A second set will help guide teachers to analyze receptive response item formats (true-false, multiple-choice, and matching items). A third set will aid with the different types of productive response item formats (fill-in, short-response, and task), and a fourth set will aid teachers in formatting personal response item formats (conferences, portfolios, self-assessments). In all cases, the purpose is to help teachers improve the formatting of the items that they use in their language tests.

Table 3.1 General guidelines for most item formats

Checklist Questions	Yes	No
1. Is the item format correctly matched to the purpose and content of the item?	☐	☐
2. Is there only one correct answer?	☐	☐
3. Is the item written at the students' level of proficiency?	☐	☐
4. Have ambiguous terms and statements been avoided?	☐	☐
5. Have negatives and double negatives been avoided?	☐	☐
6. Does the item avoid giving clues that could be used in answering other items?	☐	☐
7. Are all parts of the item on the same page?	☐	☐
8. Is only relevant information presented?	☐	☐
9. Have race, gender, and nationality bias been avoided?	☐	☐
10. Has at least one other colleague looked over the items?	☐	☐

GENERAL GUIDELINES

Table 3.1 shows some general guidelines, which are applicable to most language testing formats. They are in the form of questions that teachers can ask themselves when writing or critiquing any type of item format. In most cases, the purpose of asking these questions is to insure that the students score high or low on the item type for the right reasons. In other words, the students should answer the items correctly only if they know the concept or skill being tested or have the skill involved. By extension, the students should answer incorrectly only if they do not know the material or lack the skill being tested. Let's consider each question in Table 3.1.

1. Is the item format correctly matched to the purpose and content of the item?

Teachers will, of course, want their item formats to match the purpose and content of the item. In part, this means matching the right type of item to what is being tested in terms of modes (productive or receptive) and channels (written or oral language). For instance, teachers may want to avoid using a multiple-choice format, which is basically receptive mode (students read and select, but produce nothing), for testing productive skills like writing and speaking. Similarly, it would make little sense to require the students to read aloud (productive) the individual letters of the words in a book in order to test the receptive skill of reading comprehension. Such a task would be senseless, in part because the students would be using both receptive and productive modes mixed with both oral and written channels when the purpose of the test, reading comprehension, is essentially receptive mode and written channel. A second problem would arise because the students would be too narrowly focused in terms of content on reading the letters of the words. To avoid mixing modes and channels and to focus the content at the comprehension level of the reading skill, teachers might more profitably have the students read a written passage and use receptive-response items in the form of multiple-choice comprehension questions. In short, teachers must think about what they are trying to test in terms of all the dimensions discussed in the previous chapter and try to match their purpose with the item format that most closely resembles it.

REVIEW QUESTIONS

1. What are the theoretical and practical issues that must be considered in developing language tests? How are the theoretical issues different in general from those classified as practical?
2. On a continuum of methodological choices that ranges from structural language teaching to communicative, where would your philosophy of teaching fit? What about your philosophy of testing? Are you prescientific? Are you a psychometric-structuralist? An integrative-sociolinguist? Or are you part of the communicative wave of the future?
3. How are performance testing and task-based testing related? Different? How are they communicative in nature?
4. What is the difference between competence and performance as discussed by Chomsky? And why might this distinction be important to think about with regard to language testing?
5. What is the fundamental difference between a discrete-point test and an integrative one? Can you think of at least one example of each? Would you prefer to use discrete-point or integrative tests for purposes of placing students into the levels of a language program? Why?
6. What are the five characteristics of a communicative test? What are the three bases for rating communicative language performance? And, what are the four main components of communicative competence (according to Canale 1983b)?
7. Why is objectivity important to language testers? Under what conditions could you justify sacrificing some degree of objectivity? And why?
8. What are some of the logistical conditions that you should consider in any testing project? Which of the three logistical conditions discussed in this book (ease of construction, administration, and scoring) do you think is the most important? How are ease of test construction and ease of scoring inversely related?
9. What are the factors that you must consider in looking at the quality of a test? Which do you think are the most important?
10. What are the factors that you must keep in mind in putting together a successful testing program? Which factors do you think are the most important?
11. What is an address in a spreadsheet? What is a cell? How many columns does your spreadsheet have? How many rows?

APPLICATION EXERCISES

A. Locate a test that you think might be useful in a language program in which you are now working, or if you have never taught, find a test for an elementary, secondary, adult education, commercial, or university language program. Examine the test very carefully using Table 2.5 (p. 35), keeping in mind all the theoretical and practical issues discussed in this chapter. Perhaps you should consult with several colleagues and find out what they think of it. What differences do you now have with your colleagues in your views on testing?

B. What theoretical and practical issues would be of particular importance for implementing the test that you selected for the above application exercise (see Table 2.5, p. 35)?

DEVELOPING GOOD QUALITY LANGUAGE TEST ITEMS

INTRODUCTION

In this chapter, I will begin explaining the elements that make up a good test. The basic unit of any test is the test item, so I will begin the chapter with a broad definition of this crucial concept. Then I will continue with guidelines for item format analysis including four separate sets: general guidelines for all types of test items; guidelines for receptive response items (true-false, multiple-choice, and matching); guidelines for productive response items (fill-in, short-response, and task); and personal response items (self-assessments, conferences, and portfolios). As usual, I will end the chapter with review questions and applications exercises.

WHAT IS A TEST ITEM?

The Multilingual Glossary of Language Testing Terms (ALTE 1998, p. 149) defines an *item* as follows: "Each testing point in a test which is given a separate mark or marks." That is fine as far as it goes, but what is a "testing point" and what is a "separate mark"? I think there is a clearer way to look at test items.

In the same sense that the phoneme is a basic unit in phonology and the morpheme is a basic unit in syntax, an *item* is the basic unit of language testing. Like the linguistic units above, the item is sometimes difficult to define. Some types of items, like multiple-choice or true-false items, are relatively easy to identify because they are the individual test questions that anyone can recognize as discrete units. An item may prove more difficult to identify for the more integrative types of language tests such as dictations, interviews, role plays, and compositions, or for more personal assessments like conferences, self-assessments, or portfolios. To accommodate the variety of discrete-point, integrative, and personal item types found in language testing, I will define the term **item** very broadly as the smallest unit that produces distinctive and meaningful information or feedback on a test when it is scored or rated. This definition will be general enough to work for every type of language test from multiple-choice to portfolio, yet will be specific enough to also prove useful.

Since the *item* is the basic unit, or building block, in testing, one way to improve a test is to examine the individual items and revise the test so that only those items that are performing well remain in the revised version of the test. Teachers often look at the total scores of their students on a test, but careful examination of the individual items that contributed to the total scores can also prove very illuminating. This process of carefully inspecting individual test items is called item analysis.

More formally, **item analysis** is the systematic evaluation of the effectiveness of the individual items on a test. This is usually done for purposes of selecting the "best" items which will remain on a revised and improved version of the test. Sometimes,

2. Is there only one correct answer?

The issue of making sure that each question has only one correct answer is not as obvious as it might at first seem. Correctness is often a matter of degrees rather than an absolute. For instance, in the following item there are two possible answers (A or C) depending on how the reader sees the context:

The apple is located on _____ table.

Ⓐ a Ⓒ the

Ⓑ an Ⓓ (no article)

That problem can be corrected by clarifying the context so that only one answer will work (C):

Do you see the chair and table? The apple is on _____ table.

Ⓐ a Ⓒ the

Ⓑ an Ⓓ (no article)

Sometimes, an option that is correct to one person may be less so to another, and an option that seems incorrect to the teacher may appear to be correct to many of the students. Such differences may occur due to differing points of view on the world or to differing contexts that people can mentally supply in answering a given question. Every teacher has probably disagreed with the "correct" answer on some test that they have taken or given. Such problems arise because the item writer was unable to take into account every possible point of view. One way that test writers attempt to circumvent this problem is by having the examinees select the *best* answer. Such wording does ultimately leave the judgment as to which is the *best* answer in the hands of the test writer, but how ethical is such a stance? I feel that the *best* course of action is to try to write items for which there is clearly only one correct answer. The statistics discussed in the next chapter under ***Item Efficiency Analysis*** will help to spot cases where the results indicate that two answers are possible, or that a second answer is very close to correct.

3. Is the item written at the students' level of proficiency?

Each item should be written at approximately the level of proficiency of the students who will take the test. For instance, an item like the following (based on a reading passage not shown here) would obviously contain vocabulary that is far too difficult for most ESL students (and many native speakers of English):

According to the passage, antidisestablishmentarianism diverges fundamentally from the conventional proceedings and traditions of the Church of England.

Ⓣ Ⓕ

Since a given language program may include students with a wide range of abilities, teachers should think in terms of using items that are at about the *average* ability level for the group. To begin with, teachers may have to gauge this average level by

intuition, but later, using the item statistics provided in this chapter, they will be able to more rationally identify which items on average are too difficult, too easy, or at the appropriate level of difficulty for their students.

4. Have ambiguous terms and statements been avoided?

Ambiguous and tricky language should be avoided unless the purpose of the item is to test ambiguity. For instance, a short-answer item like the following (again based on a reading passage that does not appear here) would be ambiguous to some students:

Why are statistical studies inaccessible to language teachers in Brazil according to the reading passage?

If the correct answer was something like "language teachers get very little training in mathematics" and/or "such teachers are naturally averse to numbers," students who answered that "the libraries may be far away" would be wrong because of the ambiguity of the word *inaccessible* (even if that is factually true and mentioned in the passage).

The problem is that ambiguous language may cause students to answer incorrectly even though they know the correct answer. Such an outcome is always undesirable. Getting a colleague or two to proofread the test or having several former students take the test and comment on the items can solve this kind of problem.

5. Have negatives and double negatives been avoided?

Likewise, the use of negatives and double negatives may be needlessly confusing and should be avoided unless the purpose of the item is to test negatives. For example:

One theory that is not unassociated with Noam Chomsky is:

Ⓐ Transformational generative grammar Ⓒ Non-universal phonology

Ⓑ Case grammar Ⓓ Acoustic phonology

Clearly, the three negatives (*not*, *un-*, and *non-*) in this item make the item impossible to process. Whereas the following accomplishes the same thing without confusion (even though it contains the single negative *non-*):

One theory that is associated with Noam Chomsky is:

Ⓐ Transformational generative grammar Ⓒ Non-universal phonology

Ⓑ Case grammar Ⓓ Acoustic phonology

In those rare cases where negatives must be tested, like the example above, wise test writers use only one negative word and emphasize it (by underlining them, typing them in capital letters, or putting them in bold-faced type, as in not, NEVER, *in*consistent, etc.) so the students are sure to notice what is being tested. Students should *not* miss an item because they did *not* notice a negative marker, if indeed they know the answer.

6. Does the item avoid giving clues that could be used in answering other items?

Teachers should also avoid giving clues in one item that will help answer another item. For instance, a clear example of a grammatical structure may appear in one item that will help some students to answer a question about that structure later in the test.

Students should answer the latter item correctly only if they know the concept or skill involved, not because they were clever enough to remember and look back to an example or model of it in a previous item.

7. Are all parts of the item on the same page?

All the parts of each item should be on one page. Students, who know the concept or skill being tested, should not respond incorrectly simply because they did not realize that the correct answer was on the next page. This issue is easily checked, but sometimes forgotten.

8. Is only relevant information presented?

Teachers should also avoid including extra information that is irrelevant to the concept or skill being tested. Since most teachers will probably want their tests to be relatively efficient, any extra information not related to the material being tested should be avoided, because it will just take extra time for the students to read and will add nothing to the test. Such extra information may also inadvertently provide the students with clues they can use in answering other items.

9. Have race, gender, and nationality bias been avoided?

All teachers should also be on the alert for bias that may have crept into their test items. Race, gender, religion, nationality, age, ethnicity, and other biases must be avoided at all costs, not only because they are unethical, morally wrong, and illegal in many countries, but also because they affect the fairness and objectivity of the test. The most famous example of this was the so-called "white picket fence" item on an IQ test. This item apparently required knowledge of what a *white picket fence* is in order to answer it correctly. The item was judged biased against inner city blacks, who seldom, if ever, would see such a suburban lawn fence. The item was meant to test IQ, but instead was testing vocabulary knowledge, vocabulary that one particular group of students was unlikely to know.

The problem is that an item that is biased against one group of people is testing something in addition to what it was originally designed to test, and such an item cannot provide clear and easily interpretable information. The only practical way to avoid bias in most situations is to examine the items carefully and have other language professionals also examine them. Preferably these colleagues will be both male and female and will be drawn from different racial, religious, nationality, age, and ethnic groupings. Since the potential for bias differs from situation to situation, individual teachers will have to determine what is appropriate for avoiding bias in the items administered to their particular populations of students. Statistical techniques can help spot and avoid this bias in items too. However, these bias statistics are well beyond the scope of this book.

10. Has at least one other colleague looked over the items?

Regardless of any problems that teachers may find and correct in their items, they should always have at least one or more colleagues look over and perhaps take the test so that any additional problems may be spotted before the test is actually used to make decisions about students' lives. A related point for teachers who are not native speakers of the language being tested is the possible necessity of having native speakers take the test or at least look it over. As far back as 1961, Lado put it this way, "...if the test is administered to native speakers of the language they should make very high marks on it or we will suspect that factors other than the basic ones of language have been introduced into the items" (p. 323).

Receptive Response Items

Table 3.2 includes other questions that are specifically designed for receptive response items. **Receptive response items** require the student to select a response rather than actually produce one. In other words, the responses involve receptive language in the sense that the item responses from which students must select are heard or read, receptively. Receptive response item formats include true-false, multiple-choice, and matching items.

Table 3.2 Guidelines for receptive response items

Item Format Checklist Questions	Yes	No
True-False		
1. Is the statement worded carefully enough so it can be judged without ambiguity?	☐	☐
2. Have "absoluteness" clues been avoided?	☐	☐
Multiple-Choice		
1. Have all unintentional clues been avoided?	☐	☐
2. Are all of the distracters plausible?	☐	☐
3. Has needless redundancy been avoided in the options?	☐	☐
4. Has the ordering of the options been carefully considered? Or are the correct answers randomly assigned?	☐	☐
5. Have distracters like "none of the above," "A and B only," etc. been avoided?	☐	☐
Matching		
1. Are there more options than premises?	☐	☐
2. Are options shorter than premises to reduce reading?	☐	☐
3. Are the option and premise lists related to one central theme?	☐	☐

True-False

True-false items are typically written as statements, and students must decide whether the statements are true or false. There are two potential problems shown in Table 3.2 that teachers should consider in developing items in this format.

1. Is the statement worded carefully enough so it can be judged without ambiguity?

The statement should be carefully worded to avoid any ambiguities that might cause the students to miss it for the wrong reasons. The wording of true-false items is particularly difficult and important. Teachers are often tempted to make such items "tricky" so that the items will be difficult enough for intermediate or advanced language students. Such trickiness should be avoided: students should miss an item because they do not know the concept or have the skill being tested rather than because the item is tricky.

2. Have "absoluteness" clues been avoided?

Teachers should also avoid absoluteness clues. Absoluteness clues allow students to answer correctly without knowing the correct response. Absoluteness clues include terms like *all, always, absolutely, never, rarely, most often,* and so forth. True-false items that include such terms are very easy to answer regardless of concept or skill being tested because the answer is inevitably *false.* For example:

This book is always crystal clear in all its explanations.

(T) (F)

MULTIPLE-CHOICE

Multiple-choice items are made up of an **item stem**, or the main part of the item at the top, a **correct answer**, which is obviously the choice (usually, a., b., c., or d.) that will be counted correct, and the **distracters**, which are those choices that will be counted as incorrect. These incorrect choices are called distracters because they should distract, or divert the students' attention away from the correct answer if the students really do not know which is correct. The term **options** refers collectively to all the alternative choices presented to the students including the correct answer and the distracters. All these terms are necessary for understanding how multiple-choice items function. Five potential pitfalls for multiple-choice items appear in Table 3.2 (p. 47).

1. Have all unintentional clues been avoided?

Teachers should avoid unintentional clues (grammatical, phonological, morphological, and so forth) that help students to answer an item without having the knowledge or skill being tested. To avoid such clues, teachers should write multiple-choice items so that they clearly test only one concept or skill at a time. Consider the following item:

The fruit that Adam ate in the Bible was an _____.

(A) pear
(B) banana
(C) apple
(D) papaya

The purpose of this item is neither clear nor straightforward. If the purpose of the item is to test cultural or biblical knowledge, an unintentional grammatical clue (in that the article *an* must be followed by a noun that begins with a vowel) is interfering with that purpose. Hence, a student who knows the article system in English can answer the item correctly without ever having heard of Adam. If, on the other hand, the purpose of the item is to test knowledge of this grammatical point, why confuse the issue with the cultural/biblical reference? In short, teachers should avoid items that are not straightforward and clear in intent. Otherwise, unintentional clues may creep into their items.

2. Are all of the distracters plausible?

Teachers should also make sure that all the distracters are plausible. If one distracter is ridiculous, that distracter is not helping to test the students. Instead, those students who are guessing will be able to dismiss that distracter and improve their chances of answering the item correctly without really knowing the correct answer. An example (based again on a reading passage about Eve and Adam not shown here) follows:

Adam ate _____.

(A) an apple
(B) a banana
(C) an apricot
(D) a tire

Clearly, tire is not a plausible answer in this set. Why would any teacher write an item that has ridiculous distracters? Brown's law may help explain this phenomenon. Brown's law: when writing four-option multiple-choice items, the stem and correct option are easy to write, and the next two distracters are relatively easy to make up, as well, but the last distracter is absolutely impossible. The only way to understand Brown's law is to actually try writing a few four-option multiple-choice items. The point is that teachers are often tempted to put something ridiculous for that last distracter, simply because they are having trouble thinking of an effective distracter. Therefore, always check to see that all the distracters in a multiple-choice item are truly distracting.

3. Has needless redundancy been avoided in the options?

In order to make a test reasonably efficient, teachers should double check that items contain no needless redundancy. For example, consider the following item designed to test the past tense of the verb *to fall*:

The boy was on his way to the store, walking down the street, when he stepped on a piece of cold wet ice and _____.

(A) fell flat on his face
(B) fall flat on his face
(C) felled flat on his face
(D) falled flat on his face

In addition to the problem of providing needless words and phrases throughout the stem, the phrase "flat on his face" is repeated four times in the options, when it could just as easily have been written one time in the stem. The item could have been far shorter to read and less redundant, yet equally effective if it had been written as follows:

The boy stepped on a piece of ice and _____ flat on his face.

(A) fell
(B) fall
(C) felled
(D) falled

4. Has the ordering of the options been carefully considered? Or are the correct answers randomly assigned?

Any test writer may unconsciously introduce a pattern into the test that will help the students who are guessing to increase the probability of answering an item correctly. A teacher might decide that the correct answer for the first item should be *C* For the second item, that teacher might decide on *D* and for the third item *A*. Having already picked *C*, *D*, and *A* to be correct answers in the first three items, the teacher will very likely pick *B* to be the correct answer in the next item. Human beings seem to have a need to balance things out like this, and such patterns can be used by clever test takers to help them guess at better than chance levels without actually knowing the answer. Since testers want to maximize the likelihood that students answer items correctly because they know the concepts being tested, they generally avoid patterns that can help students guess.

A number of strategies can be used to avoid creating patterns. If the options are always ordered from the shortest to longest or alphabetically, the choice of which option is correct is out of the test writer's hands. Hence that human tendency to create patterns will be avoided. Another strategy that can be used is to randomly select which option will be correct. Selection can be done with a table of random numbers or with the aces, twos, threes, and fours taken from a deck of cards. In all cases, the purpose is to eliminate patterns that may help students guess the correct answers if they do not know them.

5. Have distracters like "none of the above," "A and B only," etc. been avoided?

Teachers can also be tempted (often due to Brown's law, mentioned above) to use options like "all of the above," "none of the above," and "*A* and *B*" I normally advise avoiding this type of option unless the specific purpose of the item is to test two things at a time and students' abilities to interpret such combinations. For the reasons discussed in Points 1 and 2 (p. 48), such items are usually inadvisable.

Matching Items

Matching items present the students with two columns of information; the students must then find and identify matches between the two sets of information. For the sake of discussion, the information given in the left-hand column will be called the **matching item premise** and that shown in the right-hand column will be labeled **options.** Thus, in a matching test, students must match the correct option to each premise. There are three guidelines that teachers should apply to matching items.

1. Are there more options than premises?

More options should be supplied than premises so that students cannot narrow down the choices as they progress through the test simply by keeping track of the options that they have already used. For example, in matching ten definitions (premises) to a list of ten vocabulary words (options), a student who knows nine will be assured of getting the tenth one correct by the process of elimination without knowing it. If, on the other hand, there are ten premises and 15 options, this problem is minimized.

2. Are options shorter than premises to reduce reading?

The options should usually be shorter than the premises because most students will read a premise then search through the options for the correct match. By controlling the length of the options as described here, the amount of reading will be minimized. Teachers often do exactly the opposite in creating vocabulary matching items by using the vocabulary words as the premises, and using the definitions (which are much longer) as the options.

3. Are the option and premise lists related to one central theme?

The premises and options should be logically related to one central theme that is obvious to the students. Mixing different themes in one set of matching items is not a good idea because it may confuse the students and cause them to miss items that they would otherwise answer correctly. For example, lining up definitions and the related vocabulary items is a good idea, but also mixing in matches between graphemic and phonemic representations of words would only cause confusion. The two different themes could be much more clearly and effectively tested as separate sets of matching items.

Productive Response Items

Table 3.3 includes additional questions that should be applied to productive response items. **Productive response items** require the students to actually produce responses rather than just select them receptively. In other words, the responses involve productive language in the sense that the answers must either be written or spoken. Productive item formats include fill-in, short-response, and task types of items.

Table 3.3 Guidelines for productive response items

Item Format Checklist Questions	Yes	No
Fill-In		
1. Is the required response concise?	☐	☐
2. Is there sufficient context to convey the intent of the questions to the students?	☐	☐
3. Are the blanks of standard length?	☐	☐
4. Does the main body of the question precede the blank?	☐	☐
5. Has a list of acceptable responses been developed?	☐	☐
Short-Response		
1. Is the item formatted so that only one relatively concise answer is possible?	☐	☐
2. Is the item framed as a clear and direct question?	☐	☐
Task		
1. Is the student's task clearly defined?	☐	☐
2. Is the task sufficiently narrow (and/or broad) for the time available?	☐	☐
3. Have scoring procedures been worked out in advance with regard to the approach that will be used?	☐	☐
4. Have scoring procedures been worked out in advance with regard to the categories of language that will be rated?	☐	☐
5. Have scoring procedures been clearly defined in terms of what each score within each category means?	☐	☐
6. Is scoring to be as anonymous as possible?	☐	☐

Fill-In Items

Fill-in items are those wherein a word or phrase is replaced by a blank in a sentence or longer text and the student's job is to fill in that missing word or phrase. There are five sets of issues that teachers should consider when using fill-in items.

1. Is the required response concise?

In answering fill-in items, students will often write alternative answers that the teacher did not anticipate when the items were written. For example in the following fill-in item there are many possible answers: John walked down the street _____. Indeed almost any adverb would work, e.g., slowly, quickly, pensively, angrily, carefully, etc.

To guard against this possibility, teachers should check to make sure that each item has one very concise correct answer. For example, a blank with only one acceptable

answer (fell) would be the following: John stepped onto the ice and immediately _____ down hard.

Alternatively, the teacher can develop a glossary of acceptable answers for each blank. Obviously, as the number of alternative possibilities rises for each item, the longer and more difficult the scoring becomes. One goal should be to create an answer key that will help make clear-cut decisions as to whether each item is correct. Another goal should be to create an answer key that is so complete that no modifications will be necessary during the scoring process because such modifications necessitate backtracking and rescoring tests that have already been scored.

2. Is there sufficient context to convey the intent of the question to the students?

In deciding how much context to provide for each blank (that is, how many words or phrases each item should contain), teachers should make sure that enough context has been provided so the purpose, or intent, of the item is clear to those students who know the answer. At the same time, avoid giving too much extra context. Extra context will burden students with extraneous material to read (see Table 3.1 no. 8) and may inadvertently provide students with extraneous clues (see Table 3.1 no. 6).

3. Are the blanks of standard length?

Generally speaking, all the blanks in a fill-in test should be the same length, that is, if the first blank is twelve spaces long, then, all the items should have blanks with twelve spaces. Blanks of uniform length do not provide extraneous clues about the relative length of the answers. Obviously, this stricture would not apply if a teacher purposely wants to indicate the length of each word or the number of words in each blank.

4. Does the main body of the question precede the blank?

Teachers should also consider putting the main body of the item before the blank in most of the items so that the students have the information necessary to answer the item when they encounter the blank. For example: Based on the above sentence, teachers should put the main body of the question before the _____. Such a strategy will help to make the test more efficient. Of course, situations do exist in language testing wherein the blank must be early in the item (for instance, when trying to test for the head noun in a sentence), but as a general rule, the blank should occur relatively late in the item.

5. Has a list of acceptable responses been developed?

In situations where the blanks may be very difficult and frustrating for the students, teachers might consider supplying a list of responses from which the students can choose in filling in the blanks. This list will not only make answering the items easier for the students, but will also make the correction of the items easier for the teacher because the students will have a limited set of possible answers to draw on. However, even a minor modification like this one can dramatically change the nature of the items. In this case, the modification would change them from productive response items to selected response items.

SHORT-RESPONSE ITEMS

Short-response items are usually items that the students can answer in a few phrases or sentences. This type of item should conform to at least the following two guidelines.

1. Is the item formatted so that only one relatively concise answer is possible?

Teachers should make sure that the item is formatted so that there is one, and only one, concise answer or set of answers that they are looking for in the responses to each item. The parameters for what will be considered an acceptable answer should be thought through carefully and clearly delineated before correcting such items. As in Point 1 for fill-in items (p. 51), the goal in short-response items is to ensure that the answer key will help the teacher make clear-cut decisions as to whether each item is correct without making modifications as the scoring progresses. Therefore, the teacher's expectations should be thought out in advance, recognizing that subjectivity may become a problem because the teacher will necessarily be making judgments about the relative quality of the students' answers. Thus, partial credit often becomes an issue with this type of item. **Partial credit** entails giving some credit for answers that are not 100 percent correct. For instance, on one short response item, a student might get two points for an answer with correct spelling and correct grammar, but only one point if either grammar or spelling were wrong, and no points if both grammar and spelling were wrong. Like all the other aspects of scoring short-response items, any partial credit scheme must be clearly thought out and delineated before scoring starts so that backtracking and rescoring will not be necessary.

2. Is the item framed as a clear and direct question?

Short-response items should generally be phrased as clear and direct questions. Unnecessary wordiness should particularly be avoided with this type of item so that the range of expected answers will stay narrow enough to be scored with relative ease and objectivity. You may even want to consider giving the students some idea of the shape of the answer you are looking for. For example (based on a reading passage about doing research not supplied here, where the expected answer given in the passage would include some form of the following three steps: gather information, analyze the information, report the results):

> According to the reading passage, what are the three steps in doing research?

Such a question would let the students know that you were looking for three things and that those things are the steps in doing research.

TASK ITEMS

Task items will be defined here as any of a group of fairly open-ended item types that require students to perform a task in the language that is being tested. A task test (or what one colleague accidentally called a *tesk*) might include a series of communicative tasks, a set of problem-solving tasks, and a writing task. In another alternative that has become increasingly popular in the last decade, students are asked to perform a series of writing tasks and revisions during a course and put them together into a portfolio (see discussion of portfolios on p. 62).

While task items are appealing to many language teachers, a number of complications may arise in using them. To avoid such difficulties, consider at least the following six guidelines.

1. Is the student's task clearly defined?

The directions for the task should be so clear that both the tester and the student know exactly what the student must do. The task may be anything that people need to

do with language. Thus, task items might require students to solve written word problems, to give oral directions on how to get to the library, to explain to another student how to draw a particular geometric shape, to write a composition on a specific topic, and so forth. The possibilities are only limited by the degree of imagination among the teachers involved. However, the point to remember is that the directions for the task must be concisely explained so the students know exactly what they are expected to do and thus cannot stray too far away from the intended purpose of the item.

2. Is the task sufficiently narrow (and/or broad) for the time available?

The task should be sufficiently narrow in scope so that it fits logistically into the time allotted for its performance. At the same time, since one purpose of task items is to get the students to produce language, the task should be broad enough so that an adequate sample of each student's language is available for proper scoring. For instance, in an essay examination, a topic that requires a yes/no answer (e.g., "Did you have a good summer?") would be far too narrow; a *wh-* question like "What did you do last summer?", though it is a cliché, is much more likely to produce a good language sample. In my high school American Literature class, I will never forget the topic assigned by my teacher for the three-hour in-class essay examination: "Explain American Literature; you have three hours." Even then, I thought that topic was way too broad for the time allowed. I wrote my heart out but only got to Emerson, missing out altogether on the chance to write about my favorite authors from the late nineteenth and twentieth centuries. In other words, I did not have enough time to adequately finish the task.

3. Have the scoring procedures been worked out in advance with regard to the approach that will be used?

Teachers must carefully work out the scoring procedures for task items for the same reasons listed in discussing the other types of productive response items. However, such planning is particularly crucial for task items because teachers have less control over the range of possible responses in such open-ended items.

Two entirely different approaches are possible in scoring tasks. A task can be scored using an **analytic approach**, in which the teachers rate various aspects of each student's language production separately; or a task can be scored using a **holistic approach**, in which the teachers use a single general scale to give a single global rating for each student's language production. The very nature of the item(s) will depend on how the teachers choose to score the task. If teachers choose to use an analytic approach, the task may have three, four, five, or even six individual bits of information, each of which should be treated as a separate item (for example, the rubric shown in Table 3.4 (p. 56) requires raters to judge five different aspects of writing: organization; logical development of ideas; grammar; punctuation, spelling, and mechanics; and style and quality of expression). A decision for a holistic approach will produce results that must be treated differently, that is, more like a single item (see Table 3.5, p. 57). Thus, teachers must decide early as to whether they will score task items using an analytic approach or a holistic one.

4. Have scoring procedures been worked out in advance with regard to the categories of language that will be rated?

If teachers decide to use an analytic approach, they must then decide which categories of language to judge in rating the students' performances. Naturally, these decisions must also occur before the scoring process begins. For example, when I was

teaching ESL at UCLA, we felt that compositions should be rated analytically, with separate scores for organization, logic, grammar, mechanics, and style as shown in Table 3.4, p. 56 (see Brown & Bailey 1984). Five categories of language were important to us, but these categories are not the only possible ones. In contrast, when I was director of the English Language Institute at the UHM, we used an analytic scale that helped us rate content, organization, vocabulary, language use, and mechanics (see Jacobs, Zinkgraf, Wormuth, Hartfiel, & Hughey 1981). Thus, the teachers at UHM preferred to rate five categories of language that are different from the five categories used at UCLA. Because such decisions were often very different from course to course and program to program, decisions about which categories of language to rate should most often rest with the teachers who are involved in the teaching process. (For an example of descriptors that are used in a *holistic* six-point scale, see ETS 1996, p. 19.)

5. Have scoring procedures been clearly defined in terms of what each score within each category means?

Having worked out the approach and categories of language to rate, it is still necessary to clearly define the points on the scales for each category. Written descriptions of the kinds of language that would be expected at each score level will help. The descriptors shown in Table 3.4 on page 56 (Brown & Bailey 1984, pp. 39–41) are examples of one way to go about delineating such language behaviors in an analytic scale. Table 3.5 (p. 57) rearranges the same descriptive information to show how it would look as a holistic scale. Regardless of the form that they take, such descriptions will help ensure that the judgments of the scorers are relatively consistent within and across categories and that the scores will be relatively easy to assign and interpret. Sometimes, training workshops will be necessary for the raters so they can agree upon the definitions within each scale and develop consistency in the ways that they assign scores (see Chapter 8 under ***Rater reliabilities***). However, as McNamara (1996, p. 26) points out, rater training may only succeed in making raters more self-consistent and may not resolve average differences in ratings between raters. He goes on to argue that such differences may be the natural state of affairs (pp. 232–239), and that, in any case, such overall differences will be moderated if raters are self-consistent and multiple raters are used.

6. Is scoring to be as anonymous as possible?

Another strategy that can help make the scoring as objective as possible is to assign the scores anonymously. A few changes in testing procedures may be necessary to ensure anonymous ratings. For instance, students may have to put their names on the back of the first page of a writing task so that the raters do not know whose test they are rating. Or, if the task is audio-taped in a face-to-face interview, teachers other than the student's teachers may have to be assigned to rate the tape without knowing who they are hearing on the cassette. Such precautions will differ from task to task and situation to situation. Since they are largely a matter of common sense, teachers can work out the details for themselves. The important thing is that teachers consider using anonymity as a way of increasing objectivity.

Table 3.4 Analytic scale for rating composition tasks

	20-18 Excellent to Good	17-15 Good to Adequate	14-12 Adequate to Fair	11-6 Unacceptable	5-1 Not College-level Work
I. Organization: Introduction, Body & Conclusion	Appropriate title, effective introductory paragraph, topic is stated, leads to body; transitional expressions used; arrangement of material shows plan (could be outlined by reader); supporting evidence given for generalizations; conclusion logical & complete	Adequate title, introduction, & conclusion; body of essay is acceptable but some evidence may be lacking, some ideas aren't fully developed; sequence is logical but transitional expressions may be absent or misused	Mediocre or scant introduction or conclusion; problems with the order of ideas in body; the generalizations may not be fully supported by the evidence given; problems of organization interfere	Shaky or minimally recognizable introduction; organization can barely be seen; severe problems with ordering of ideas; lack of supporting evidence; conclusion weak or illogical; inadequate effort at organization	Absence of introduction or conclusion; no apparent organization of body; severe lack of supporting evidence; writer has not made any effort to organize the composition (could not be outlined by reader)
II. Logical Development of Ideas: Content	Essay addresses the assigned topic; the ideas are concrete and thoroughly developed; no extraneous material; essay reflects thought	Essay addresses the issues but misses some points; ideas could be more fully developed; some extraneous material is present	Development of ideas not complete or essay is somewhat off the topic; paragraphs aren't divided exactly right	Ideas incomplete; essay does not reflect careful thinking or was hurriedly written; inadequate effort in area of content	Essay is completely inadequate and does not reflect college-level work; no apparent effort to consider the topic carefully
III. Grammar	Native-like fluency in English grammar; correct use of relative clauses, prepositions, modals, articles, verb forms, and tense sequencing; no fragments or run-on sentences	Advanced proficiency in English grammar; some grammar problems don't influence communication, although the reader is aware of them; no fragments or run-on sentences	Ideas getting through to the reader, but grammar problems are apparent and have a negative effect on communication; run-on sentences or fragments present	Numerous serious grammar problems interfere with communication of the writer's ideas; grammar review of some areas clearly needed; difficult to read sentences	Severe grammar problems interfere greatly with the message; reader can't understand what the writer is trying to say; unintelligible sentence structure
IV. Punctuation, Spelling, & Mechanics	Correct use of English writing conventions; left & right margins, all needed capitals, paragraphs indented, punctuation & spelling; very neat	Some problems with writing conventions or punctuation; occasional spelling errors; left margin correct; paper is neat and legible	Uses general writing conventions but has errors; spelling problems distract reader; punctuation errors interfere with ideas	Serious problems with format of paper; parts of essay not legible; errors in sentence-final punctuation; unacceptable to educated readers	Complete disregard for English writing conventions; paper illegible; obvious capitals missing, no margins, severe spelling problems
IV. Style & Quality of Expression	Precise vocabulary usage; use of parallel structures; concise; register good	Attempts variety; good vocabulary; not wordy; register OK; style fairly concise	Some vocabulary misused; lacks awareness of register; may be too wordy	Poor expression of ideas; problems in vocabulary; lacks variety of structure	Inappropriate use of vocabulary; no concept of register or sentence variety

Table 3.5 Holistic version of the scale for rating composition tasks

Scores	Descriptors
5	Appropriate title, effective introductory paragraph, topic is stated, leads to body; transitional expressions used; arrangement of material shows plan (could be outlined by reader); supporting evidence given for generalizations; conclusion logical & complete. Essay addresses the assigned topic; the ideas are concrete and thoroughly developed; no extraneous material; essay reflects thought. Native-like fluency in English grammar; correct use of relative clauses, prepositions, modals, articles, verb forms, and tense sequencing; no fragments or run-on sentences. Correct use of English writing conventions; left & right margins, all needed capitals, paragraphs indented, punctuation & spelling; very neat. Precise vocabulary usage; use of parallel structures; concise; register good.
4	Adequate title, introduction, & conclusion; body of essay is acceptable but some evidence may be lacking, some ideas aren't fully developed; sequence is logical but transitional expressions may be absent or misused. Essay addresses the issues but misses some points; ideas could be more fully developed; some extraneous material is present. Advanced proficiency in English grammar; some grammar problems don't influence communication, although the reader is aware of them; no fragments or run-on sentences. Some problems with writing conventions or punctuation; occasional spelling errors; left margin correct; paper is neat and legible. Attempts variety; good vocabulary; not wordy; register OK; style fairly concise.
3	Mediocre or scant introduction or conclusion; problems with the order of ideas in body; the generalizations may not be fully supported by the evidence given; problems of organization interfere. Development of ideas not complete or essay is somewhat off the topic; paragraphs aren't divided exactly right. Ideas getting through to the reader, but grammar problems are apparent and have a negative effect on communication; run-on sentences or fragments present. Uses general writing conventions but has errors; spelling problems distract reader; punctuation errors interfere with ideas. Some vocabulary misused; lacks awareness of register; may be too wordy.
2	Shaky or minimally recognizable introduction; organization can barely be seen; severe problems with ordering of ideas; lack of supporting evidence; conclusion weak or illogical; inadequate effort at organization. Ideas incomplete; essay does not reflect careful thinking or was hurriedly written; inadequate effort in area of content. Numerous serious grammar problems interfere with communication of the writer's ideas; grammar review of some areas clearly needed; difficult to read sentences. Serious problems with format of paper; parts of essay not legible; errors in sentence-final punctuation; unacceptable to educated readers. Poor expression of ideas; problems in vocabulary; lacks variety of structure.
1	Absence of introduction or conclusion; no apparent organization of body; severe lack of supporting evidence; writer has not made any effort to organize the composition (could not be outlined by reader). Essay is completely inadequate and does not reflect college-level work; no apparent effort to consider the topic carefully. Severe grammar problems interfere greatly with the message; reader can't understand what the writer is trying to say; unintelligible sentence structure. Complete disregard for English writing conventions; paper illegible; obvious capitals missing, no margins, severe spelling problems. Inappropriate use of vocabulary; no concept of register or sentence variety.

Personal Response Items

Table 3.6 (adapted from Brown & Hudson 2002) includes additional questions that should be applied to personal response items. **Personal response items** encourage the students to produce responses that hold personal meaning. In other words, the responses allow students to communicate in ways and about things that are interesting to them personally. Personal response item formats include self-assessments, conferences, and portfolios. (For more on this general class of item formats see Bailey 1998; Genesee & Upshur 1996; O'Malley & Valdez Pierce 1996.)

Table 3.6 Guidelines for personal response items

Item Format Checklist Questions	Yes	No
Self-Assessments		
1. Have you decided on a scoring type (holistic or analytic)?	☐	☐
2. Have you decided in advance what aspect of the students' language performance they will be assessing?	☐	☐
3. Have you developed a written rating scale for the learners to use in scoring?	☐	☐
4. Does the rating scale describe concrete language and behaviors in simple terms?	☐	☐
5. Have you planned the logistics of how the students will score themselves?	☐	☐
6. Have you checked to see if students understand the self-scoring procedures?	☐	☐
7. Have you considered having another student and/or the teacher do the same scoring?	☐	☐
Conferences		
1. Have you introduced and explained conferences to the students?	☐	☐
2. Have you given the students the sense that they are in control of the conference?	☐	☐
3. Have you focused the discussion on the students' views about the learning process?	☐	☐
4. Have you considered working with students on self-image issues?	☐	☐
5. Have you elicited performances on specific skills that need to be reviewed?	☐	☐
6. Are the conferences frequently scheduled at regular intervals?	☐	☐
7. Have you scored conferences by applying Numbers 1-6 under ***Task*** in Table 3.3 (p. 51)?	☐	☐
Portfolios		
1. Have you introduced and explained portfolios to the students?	☐	☐
2. Have you and the students decided who will take responsibility for what?	☐	☐
3. Have students selected and collected *meaningful* work?	☐	☐
4. Have students periodically reflected in writing on their portfolios?	☐	☐
5. Have other students, teachers, outsiders, etc. periodically examined the portfolios?	☐	☐
6. Have you scored the portfolios by applying Numbers 1-6 under ***Task*** in Table 3.3 (p. 51)?	☐	☐

Self-Assessments

Self-assessments will be defined here as any items wherein students are asked to rate their own knowledge, skills, or performances. Thus, self-assessments provide the teacher with some idea of how the students view their own language abilities and development. The related concept of **peer assessments** is simply a variation on this theme that requires students to rate each other (see Brown 1998; Gardner 1996; McNamara & Deane 1995; Murphey 1995; Oscarson 1997).

1. Have you decided on a scoring type (holistic or analytic)?

As with the task items in the productive response section, when using self- and peer-assessments, you will need to consider whether you want to use holistic or analytic scoring. In other words, do you want the students to make their judgments holistically (i.e., making a single "gut reaction" judgment on say a scale of one to ten about the students' language performance) or analytically (i.e., making several more-detailed judgments each on its own scale of say one to five about the students' language performance). (For more on this distinction, see Number 3 in the *Task* section of Table 3.3, p. 51).

2. Have you decided in advance what aspect of the students' language performance they will be assessing?

If an analytic approach is to be used, you should next decide what aspects of their language performance the students will be assessing. Since this is an opportunity for the teacher to focus students' attention on particular aspects of the language, these aspects of the language performance should be selected and defined with some care. For example, when developing a self-assessment instrument for students to rate their own video-taped role plays, you could have them rate their fluency, grammar, pronunciation, vocabulary usage, pragmatics, cohesion, repair strategies, turn-taking strategies, error self-correction, body language, facial expressions, hands, etc. There are many possible categories, so the key would seem to be to select those that the teacher thinks are most germane to what and how the students are learning in that particular course. One way to do this would be to discuss the possibilities with the students and decide together (perhaps with considerable guidance from the teacher) which categories should be used and how they should be defined.

3. Have you developed a written rating scale for the learners to use in scoring?

Next, it is important to provide a written rating scale to help guide the students in their ratings. Johnson (1998) shows a scale he used for peer-assessments of speech presentations in his classes in Japan. That scale is shown in Table 3.7 (p. 60). Though simple in form, Johnson's scale would nevertheless be considered analytic because it asks the students to make separate judgments of several subcategories of language performance within the two broader categories of *voice* and *body*. This is a simple, yet effective, written rating scale. If Johnson had decided to use a holistic rating scale instead of an analytic scale, he might, for instance, have asked the students to simply make a single overall judgment of the student performance: was it (1) poor, (2) fair, (3) good, (4) great, or (5) excellent in terms of voice (volume, rate, pitch, & enunciation) and body (posture, gestures, & eye contact)?

Table 3.7 Peer-assessment rating scale

Evaluator: Presenter: _____ Title: _____

	Rating					
Skill	**Poor**	**Fair**	**Good**	**Great**	**Excellent**	**Comments**
VOICE						
Volume	1	2	3	4	5	
Rate	1	2	3	4	5	
Pitch	1	2	3	4	5	
Enunciation	1	2	3	4	5	
BODY						
Posture	1	2	3	4	5	
Gesture	1	2	3	4	5	
Eye-contact	1	2	3	4	5	

General comments:

4. Does the rating scale describe concrete language and behaviors in simple terms?

You will probably want to write the rating scale so it uses concrete language and describes the expected behaviors in simple terms. In many rating rubrics (like the analytic scale from Brown and Bailey shown in Table 3.4, p. 56, or the revised holistic version of that same scale shown in Table 3.5, p. 57), much more detail is given than that given in Table 3.7 above. One goal in any scale should be to describe in the most concrete terms possible, the language and behaviors that the students are to rate at each possible level of performance. That would be an argument for the sort of detailed descriptors found in the scales in Tables 3.4 and 3.5. However, since one goal of such self-assessment or peer-assessment scales is also to explain to students in the clearest possible terms what they should judge, arguments can also be made for the sort of simple, straightforward scale in Table 3.7 above. Both strategies have points to recommend them. You will probably want to decide which way you want to proceed based on the conditions and circumstances in your particular teaching situation.

5. Have you planned the logistics of how the students will score themselves?

It is also probably wise to work out the logistics of the self- or peer-assessments in advance. Who will rate each language performance? How will the students make their judgments? How will those judgments be recorded? How will they be collected? Compiled? Analyzed? And how will they be reported to the students who are being rated? Also, who will do all of the above? These are all questions that should be addressed in advance so that chaos does not ensue during the actual self- or peer-assessment process. As a general rule, the wisest strategy might be for the teacher to involve the students in these responsibilities as much as possible, so that students can not only learn as much as possible from the process, but also so the teacher's roles (and workload) are minimized.

6. Have you checked to see if students understand the self-scoring system?

Naturally, a clear explanation of the self-scoring (or peer-scoring) process and how it will proceed will help the students understand what they are to do and why they are doing it. Therefore, directions and explanations should be developed to include all the aspects of the assessment process explained in Questions 1–5 above.

However, explaining all of the above is not enough. You will also probably want to check to see if students understand the self-scoring procedures, and if necessary, repeat portions of your explanation or explain it more clearly.

7. Have you considered having another student and/or the teacher do the same scoring?

As discussed in our last point, teachers and students have often communicated to me that self- and peer-assessments are fine, but the students often want the teacher to do the same scoring as well. In addition, considerations of reliability (see Chapter 8) indicate that it is probably better to involve more than one student in each rating, which supports the wisdom of having another student and/or the teacher do the scoring in addition to a single self- or peer-assessment.

CONFERENCES

Conferences are defined here as any assessment procedures that involve students visiting the teacher's office alone or in groups for brief meetings. In such conferences, the teacher can assess students' abilities to perform particular language points and/or give students feedback on their work (see O'Malley & Valdez Pierce 1996; Genessee & Upshur 1996; Brown 1998).

1. Have you introduced and explained conferences to the students?

At the outset, you will want to introduce and explain the purpose of the conferences to the students. As pointed out above, those purposes may include assessing students' abilities to perform particular language points or giving students feedback on their work; as you will see below, purposes may also include discussion of students' views of the learning processes, bolstering the students' self-images, reviewing specific language skills, and so forth. Whatever the purposes, you will probably want to clearly explain them in class before setting up the appointments so students don't fear that they have done something wrong or are being singled out for punishment. In order to adequately explain the purpose of the conferences, you will probably need to explain some aspects of the following points.

2. Have you given the students the sense that they are in control of the conference?

You may find it wise to negotiate the purposes of the conferences with the students so they have a sense of control over what will be covered or discussed and how the conference will proceed. The teacher can still guide the students into working on the areas described in Questions 3, 4, and 5 below, but in doing so, advocates of this assessment procedure stress the importance of giving the students the sense that they are in control of the conference.

3. Have you focused the discussion on the students' views about the learning process?

In the process of conducting conferences, you might want to consider focusing the students' attention on their views of the language-learning processes. They may never have thought explicitly about these processes. Hence, conferences give the teacher a

chance to encourage students to reflect on what it means to learn a language and on the strategies that work best for them.

4. Have you considered working with students on self-image issues?

Another point often mentioned in the literature on conferences is that they afford the teacher an opportunity to work with students on self-confidence and self-image issues. This is particularly useful for students who lack confidence or have poor self-images when they are in the larger group of the classroom.

5. Have you elicited performances on specific skills that need to be reviewed?

From a language learning point of view, conferences afford the teacher an opportunity to elicit and work on specific language skills. You may want to try to observe which students are having trouble with which language skills in class; then, elicit and work on those skills only with the students who need it. Or it may make more sense to check everyone for the ability to perform certain skills during the conference and then work on it only for those students who need to improve in that particular area.

6. Are the conferences frequently scheduled at regular intervals?

Regardless of what you decide to do in the conferences, they should be held frequently and at regular intervals (say once every week or two). The point is that conferences are not likely to be taken seriously by the students, nor will they do much good, if they are not a regular part of the course curriculum.

7. Have you scored conferences by applying Numbers 1–6 under *Task* in Table 3.3?

Grading conferences may also persuade the students to take the conferences seriously. One way to grade conferences would be to work out (perhaps with the students) a scoring system for the conference. Another way would be to ask the students to reflect in writing on what happened during the conference and then score that. In either case, the principles described in Numbers 3–5 under *Task* in Table 3.3 (p. 51) and the associated prose will be helpful.

PORTFOLIOS

Portfolios are any procedures that require students to collect samples of their second language use (e.g., compositions, audio recording, video clips, etc.) into a box or folder for examination at some time in the future by peers, parents, outsiders, etc. Portfolios were originally developed for professional architects, painters, photographers, dancers, actors, etc. to use as examples of their work to show to prospective employers. However, portfolios have recently been adapted for educational purposes, and specifically for language-learning situations. (For more on portfolios, see Popham 1995; O'Malley & Valdez Pierce 1996; Norris 1996; Genessee & Upshur 1996; Brown 1998.)

1. Have you introduced and explained portfolios to the students?

As with conferences, at the outset, you will want to introduce and explain the purpose of the portfolios to the students. In order to do so, you will probably need to explain Questions 2–6 below.

2. Have you and the students decided who will take responsibility for what?

It is wise to work out who will be responsible for each aspect of the process of assembling the portfolios. Who will organize and keep track of the portfolios? Where will they be stored? Who will collect them and pass them out when students are to work on them? These are all questions that should be addressed in advance so that chaos does not ensue during the portfolio development process. As with the self-assessment procedures, the wisest strategy is for the teacher to involve the students in these responsibilities as much as possible, not only so that students can learn as much as possible from the process, but also so the teacher's roles and workload are minimized.

3. Have students selected and collected meaningful work?

The work that the students collect together into the portfolios should be meaningful to them. This can be accomplished by allowing them to make the selection decisions, at least to some degree. That way they can decide what is meaningful or not to them. For example, if the students will be writing nine compositions during the semester, you might negotiate with them and decide together that they will select one composition from the first set of three, one from the second set of three, and one from the last set of three (including all associated rough, second, and final drafts) to include in their portfolios so their progress in writing ability will be displayed. It may also be a good idea to encourage students to add illustrations, collages, photos, etc. to make the work more personal and meaningful to them.

4. Have students periodically reflected in writing on their portfolios?

Another important component of the portfolio development process is to have students periodically reflect in writing on their portfolios. They might reflect on how much progress they have made in their writing abilities, what they still need to work on, how their attitudes toward writing have changed in the process of developing the portfolio, etc. These reflections need not be lengthy, but they should probably be done on a regular basis, and they should be included in the portfolio.

5. Have other students, teachers, outsiders, etc., periodically examined the portfolios?

Yet another aspect of the portfolio process that is often mentioned in the literature is the importance of having other students, teachers, outsiders, etc. periodically examine the portfolios. Such examination of portfolios can be done at an open house or simply by arranging for classes that are doing portfolios to visit each other and have a look at what the members of the other class did in their portfolios. The purposes for displaying the portfolios in this way are to encourage students to take pride in them, to help students feel ownership in their work, and to make the whole process more meaningful to the students.

6. Have you scored the portfolios by applying Numbers 1–6 under *Task* in Table 3.3?

Grading portfolios may also encourage the students to take them more seriously. The principles described in Numbers 3–5 under *Task* in Table 3.3 (p. 51) and the associated prose will be helpful in setting up a holistic or analytic scoring grid for portfolios. You might find it useful to work out the scoring grid in discussions with the students.

WHY BOTHER WITH ITEM FORMAT ANALYSIS?

In short, item format analysis involves asking those questions in Tables 3.1, 3.2, 3.3, and 3.6, which are appropriate for a specific set of items and making sure that the items conform to the guidelines insofar as they apply to the particular teaching situation. Clearly, this type of item analysis relies heavily on common sense. Nevertheless, item format analysis is important because an item that is badly constructed is not likely to be effective or fair, even if the item looks like it is testing the appropriate content. In other words, good format would seem to be a precondition for effective testing of any content.

REVIEW QUESTIONS

1. What is an item?
2. What is the difference between an item and a test?
3. What is an item on a cloze test? A dictation? A composition?
4. What is item format analysis?
5. Why is item format analysis important?
6. What are the basic differences in item format analysis among receptive response, productive response, and personal response items?

APPLICATION EXERCISES

A. Find a language test that you are now using or have previously used and apply the item format analysis techniques covered in this book to critique the quality of the items on that test.

B. Read the satirical test on page 65. There are gross problems with these items; that is what makes readers laugh. Pick out as many of the violations of the item writing guidelines in this chapter as you can find and jot them down.

ESL/EFL TEACHER CERTIFICATION TEST

James Dean Brown
University of Hawai'i at Manoa

MULTIPLE-CHOICE SECTION (Time limit: one month) (2 points each):

1. What dialect of American English is spoken by people in New England?
a. Southern **b.** Midwestern **c.** Hawaiian Creole **d.** New England

2. Communicative language teaching would be best described as an _____.
a. technique **c.** type of syllabus
b. method **d.** approach

3. Where do foreign students come from?
a. supermarkets **c.** toy stores
b. drug stores **d.** other countries
(please select only one).

4. Caleb Gattegno's name is associated with _____.
a. Suggestopedia **c.** Counseling Language Learning
b. Total Physical Response **d.** THE SILENT WAY

SHORT-ANSWER SECTION (Time limit: yes) (1.875 points each):

5. In the space provided, outline the important characteristics of all major language-teaching methodologies with particular reference to grammar-translation, structuralism, audio-lingual approach, communicative language teaching, and the task-based approach, or give Gattegno's first name.

6. Can you explain transformational-generative grammar—Yes or No?

7. True or false - Morphemes are an important class of pain killers.

8. What color is Mike Long's black box?

9. What European language is spoken in French Guyana?

10. Spell Gattegno, Lozanov, Rassias, and Asher.

11. Explain Krashen's monitor model in detail—or spell your name in block letters.

ESSAY SECTION (Time limit: one hour.) (78.875 points):
Explain the history of the English language including its Indo-European, Germanic, and Latin origins. Focus primarily, but not exclusively, on the Great Vowel Shift, Grimm's Law, and the knowledge gained from works like *Beowulf* and Chaucer's *Canterbury Tales*. Be sure to list all words introduced from French since 1066. Also discuss the development of each of the British dialects (which are mostly of historical interest), as well as modern English dialects like American, Australian, Canadian, etc. (Use back of sheet if needed.)