The CATESOL Journal

JEAN L. TURNER
Monterey Institute of International Studies

Creating Content-Based Language Tests: Guidelines for Teachers

The problems that language teachers face in developing their classroom tests are especially complex in content-based programs. The eight-stage guidelines for test development presented here outline the steps that test writers should follow to create appropriate, content- and context-specific tests. A broader benefit of the guidelines is that student progress in different classes and programs can be compared with reference to how the guideline activities were completed. This allows language educators to address important issues such as the instructional value of various content areas and the overall effectiveness of a particular CBI program in comparison to other CBI programs or different types of language instruction.

The responsibility for developing tests to measure students' progress in their ESL classes usually falls to their teachers. Commercial tests, such as those that accompany textbooks, are occasionally available and appropriate, but often ESL teachers find themselves alone on a dark and dreary night, writing tests to be given the following day. This is a frustrating task; the other demands of teaching often seem much more urgent, and few teachers have received training in writing tests. In content-based language instruction (CBI), where the characteristics of the content and the content instruction determine to some extent the nature of the language instruction, developing suitable tests of student progress can be even more frustrating and complex. For example, teachers doing theme-based language instruction find that they must create a new test for each topic. The tests a teacher creates for a class centered on a particular current event, such as the reunification of Germany, are not going to work for classes that are centered on different issues. In sheltered and adjunct language programs, in which the content is taught by a content expert rather than the language teacher, there are even greater demands on the teacher developing the language tests.

The test-development guidelines presented here serve several purposes. Their most immediate purpose is to outline the relatively simple steps that test writers should follow to create consistent tests

that truly measure the extent to which students learned what they were taught. However, careful execution of the outlined steps produces more than consistent, appropriate tests; the test-development process also promotes more integrated, effective instruction because the guideline activities require the language teacher to consider both language and content objectives, or—in the case in which language and content teachers work together—they require the cooperation of the team members to clarify the purposes of their CBI program. A broader benefit is that the results of tests developed for different classes or programs can be compared with reference to how the guideline activities were executed. This allows language educators to begin to form answers to important questions such as which content areas and classes lend themselves most effectively to CBI programs. It also allows teachers to make judgments regarding the effectiveness of a particular CBI program compared to other CBI programs or other types of language instruction.

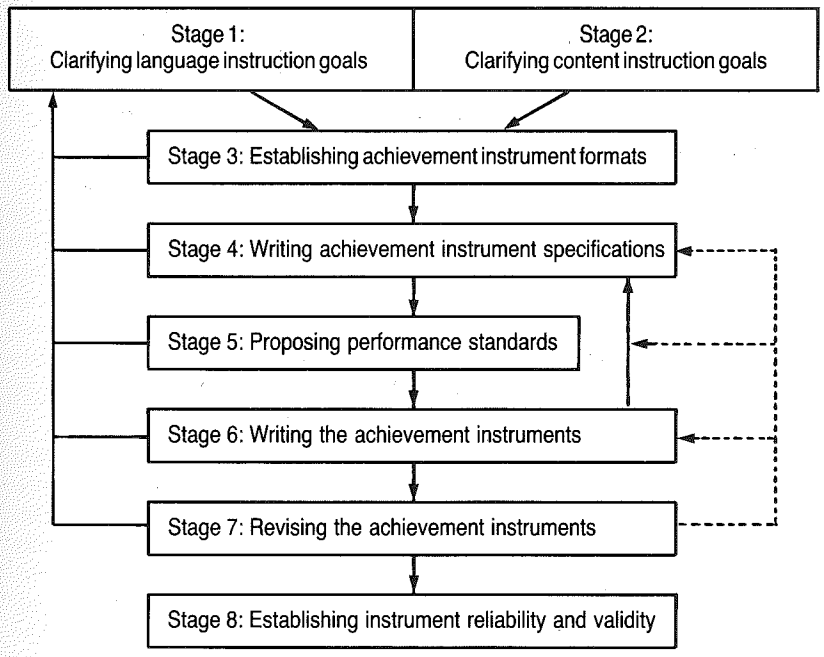
Overview of the Guidelines

The guidelines are a condensed version of the context-adaptive model for developing language achievement tests for CBI language programs (Turner, 1991). The model and the guidelines are adaptive in the sense that the manner in which the stages are completed and the nature of the tests that are written are determined by the characteristics of the class or program for which the tests are developed. The guidelines reflect sheltered- and adjunct-model CBI designs but can be used to guide the development of tests for use in theme-based programs as well.

The eight stages and the iterative nature of the test-writing process are summarized in Figure 1. The proximity of Stages 1 and 2 in the figure represent the high degree of cooperation that is required in CBI programs in which there are both language and content experts. The solid lines and arrows connecting Stages 3, 4, 5, 6, and 7 allow, if necessary, a return to Stage 1 for clarification of the instructional purposes of a program and repetition of stages which follow. The dotted lines and arrows indicate that revision of a test includes revision of the specifications, and possibly, revision of the performance standards. A detailed, illustrative discussion of each stage follows.

Figure 1.

The Adaptive Model for Developing Measures of Language Achievement

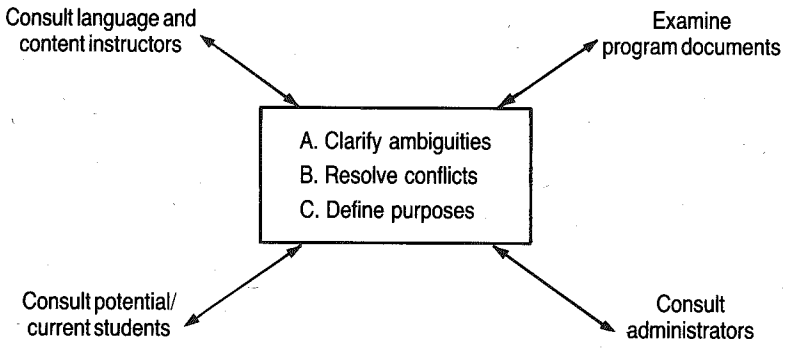


The Eight Test-Development Stages

Because tests created through this process are based on specifications derived from the instructional purposes of a particular class, it is critical that the purposes of the language and content components be clear. It is also critical that the purposes be understood and agreed upon by all participants. Stages 1 and 2, summarized in Figure 2, guide clarification of the instructional purposes. These stages may initially seem unnecessary to teachers/test writers—they already know what they want their students to learn. However, other participants in the program might have different notions of the instructional purposes. The procedures included in these two stages provide an important check of these various perspectives, revealing misunderstandings or ambiguities that should be resolved. The procedures also establish a channel of communication among the information sources, allowing negotiation of a consensus regarding the instructional goals. The two-directional arrows in Figure 2 represent this interactional quality of the guidelines.

Figure 2.

A Diagram of Stages 1 and 2 Procedures Clarifying Instructional Purposes



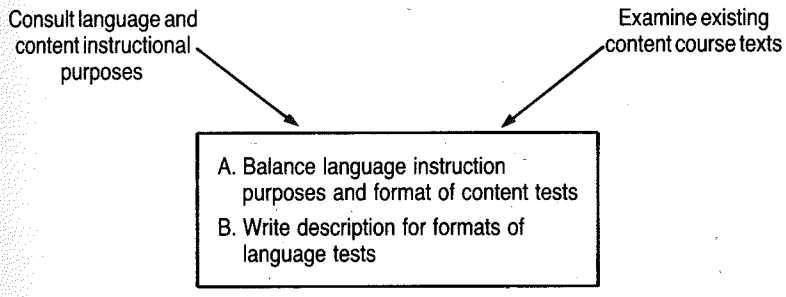
The teachers in a CBI program are perhaps the most important source of information regarding the instructional purposes of a program and should certainly be consulted to resolve any discrepancies among the information sources. As shown in Figure 2, other sources include program documentation, such as the curriculum, class descriptions, instructional materials, and existing tests. Program administrators are consulted to confirm their understanding of the purpose of the program. Students' impressions of the instructional purposes also represent an integral component in the process of clarifying the purposes and negotiating a consensus. When students' understanding of the purpose of a CBI program is different from that of the teachers, problems arise. Consider the frustration and confusion that would develop among students who believed they were studying to improve their conversational skills when the course tests reflected their teachers' belief that the purpose was to improve academic reading and writing. Reaching a consensus regarding the instructional purposes requires an exchange of information among the various sources and often results in some sort of adjustment in one or more of them. On a program level, this might involve teacher training, student orientation, or modification of program documents.

Stage 3 guidelines direct the teacher/test writer's decisions regarding what the tests will look like; for example, they might involve multiple choice items, writing an essay, or less traditional tasks such as structured story telling or problem solving. To complete Stage 3, the teacher/test writer compares the clarified instructional purposes for the language and content components and reviews any content tests. (This process is simpler in theme-based CBI programs in which the content is usually taught by the language teacher.) Stage 3 is especially important in sheltered- and adjunct-model programs because the premises on which these CBI approaches are based include the notion that language instruction should reflect the eventual uses the learner will have for the language. Students in these kinds of classes have immediate use for language; thus, it makes sense for the language tests to mirror, to whatever extent possible, the format of the tests used in the content class.

The teacher/test writer must keep in mind, however, that the format of the content class tests cannot simply be copied over into the tests for the language class. For example, if the focus of instruction for a particular adjunct language class is improvement of expository writing and the exams for the adjunct content class are multiple choice, it makes no sense to write multiple choice language tests. Instead, the situation calls for language tests which require the students to demonstrate their improved ability to produce expository writing. Figure 3 summarizes the procedures that teacher/test writers perform to define the best formats for their CBI language tests.

Figure 3.

A Diagram of Stage 3 Procedures Establishing the Test Formats



At Stage 4, test plans (specifications) for the language tests are prepared. Writing specifications involves a little more work for the teacher/test writer than simply writing tests, but having specifications to serve as a guide can help a test writer stay on track when writing tests. Specifications act as blueprints; having them means that the teacher/test writer does not have to invent or reinvent each test activity, but can simply refer to the carefully developed, clearly articulated plan. Specifications are also useful because they can be used more than once; for example, they might be used to guide the development of additional forms of a particular test.

Specifications have four components (Popham, 1978, 1981):

(a) a *general description* of the skill(s) that the test will measure;

(b) a *passage description* that shows what the text or passage that the questions are based on will look like;

(c) an *example question* or *example task* that shows what the test questions will look like and how the students will answer; and

(d) a *scoring procedure description* that specifies the characteristics of acceptable and unacceptable responses.

The test specifications developed by Macdonald (1991) to determine whether ESL students were ready to participate in a sheltered high school science class are presented below to demonstrate what these four components might actually look like. (See Appendix for the complete test developed by Macdonald.)

1. *General description:* The purpose of this test is diagnostic, that is, to determine if students are capable of participating in the sheltered science class. It measures the students' ability to read and write. It is based on observation of activities that are conducted in the sheltered science class.

- The student should be able to read the passage and demonstrate recognition of the main ideas.
- The student should demonstrate the ability to apply the main ideas to information not specifically given in the text.
- The student should be able to understand vocabulary from the context.
- The student should be able to write a one-paragraph essay that is organized, addresses the topic given, and follows basic rules of capitalization and punctuation, and, although it may contain some errors, they should not interfere with meaning.

2. *Passage description:* (The source for the test passage is a science lesson presented by the science teacher.) The criteria for the passage are as follows:

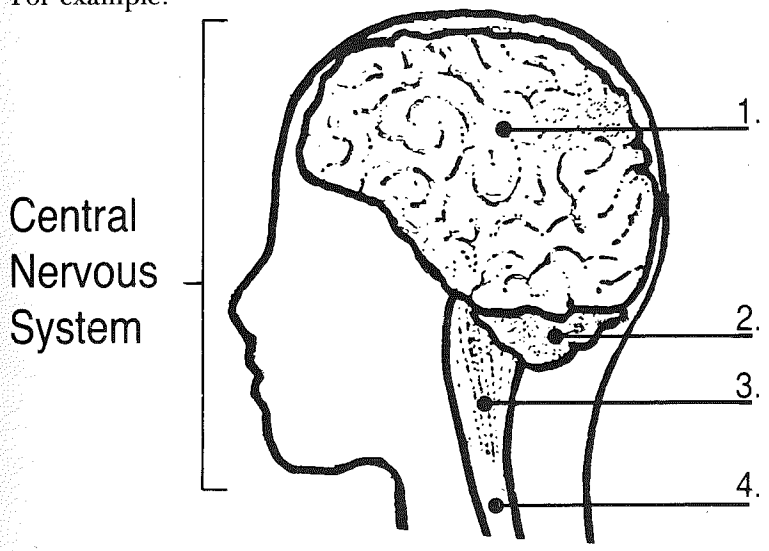
- The passage should contain all of the information that the student needs in order to complete the test, even if the student has no previous knowledge of the topic of the passage.
- The topic of the passage should be a topic that ESL students actually study in the sheltered science class.
- The passage may contain detailed, scientific information, but this information should be explained and paraphrased using terms that the students are likely to understand.

3. *Example questions and tasks:* (There are four types of items on this test.)

Type 1: Labeling

Using the diagram below, label the four major parts of the central nervous system.

For example:



Type 2: Matching

Draw a line from each part of the central nervous system to the activities that it controls.

For example:

Central Nervous System Part	Activity
Example:	Walking
Cerebrum	

Type 3: Vocabulary

Complete the following sentences using the most correct vocabulary word from this list.

For example:

coordination	involuntary
memory	to control
paralyzed	

(a) In order to play sports, you need good _____ .

Type 4: Essay

For example:

Write a one paragraph essay explaining what parts of the brain are most important when you a playing a sport. You may choose any sport—soccer, tennis, swimming, basketball, football, and so forth.

4. *Scoring Procedure Descriptions:*

Type 1 items: Objectively scored (right or wrong) based on an answer key.

Type 2 items: Objectively scored (right or wrong) based on an answer key.

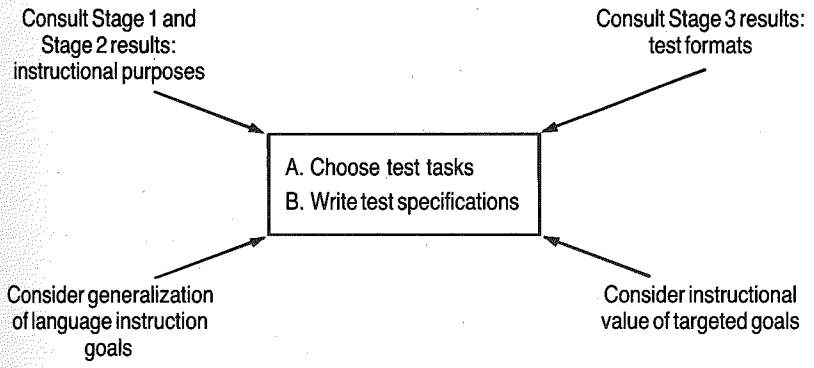
Type 3 items: Objectively scored (right or wrong) based on an answer key. Spelling and word form must be accurate to be considered correct.

Type 4 items: Subjectively scored using a holistic approach. The essays should be read twice and rated holistically for: grammar (5 points); vocabulary (5 points); mechanics (5 points); and content (5 points). Students receive one point for following the instructions and attempting to respond to the essay.

Like all tests, careful review might reveal areas that could be improved; thus Macdonald's test specifications (and test) are included here not as a model for developing CBI tests, but rather as an example of how an individual teacher/test writer applied the test-development methodology presented here to create an appropriate, context-specific test. A teacher/test writer developing tests for a different type of content-based class or a different purpose (e.g., an achievement test vs. a diagnostic test) might create tests quite different from the test developed by Macdonald. Figure 4 summarizes the main steps a teacher/test writer should follow in writing specifications for a test.

Figure 4.

A Diagram of Stage 4 Procedures
Writing the Test Specifications



As indicated in Figure 4, the clarified instructional purposes for the language and content components of a given program should be held in mind when writing test specifications. In addition, the teacher/test writer must consider the generalizability of potential test tasks. For example, if a teacher wanted to measure students' improvement in expository writing, measuring their ability to write isolated sentences would be inadequate. Although the formation of individual sentences is a component of expository writing, it cannot be assumed that students who write acceptable sentences can also write acceptable paragraphs.

In addition to the generalizability of tasks, the teacher/test writer must also consider the instructional value of tasks (Popham, 1981). Test plans should specify tasks that both the teacher and the students understand and perceive to be important. It is also critical that the teacher and students understand and agree upon the characteristics of successful accomplishment of the tasks. The students should know what successful completion of the tasks looks like (or sounds like) even if they are not yet able to produce acceptable renditions. When writing test specifications, both the generalizability and instructional value of potential tasks are weighed with the results of Stage 3 in mind, in which the format of the tests is determined.

At Stage 5, how students' test performance will be interpreted is decided. This is called proposing or setting a performance standard. The procedures at Stage 5 help the teacher/test writer answer questions such as:

1. When tests yield numerical scores, what do particular scores mean; for example, is 85% correct a passing grade?

2. When letter grades are awarded, what is the correspondence between numerical scores and the letter grades that are given—is 85% an *A*, a *B* or a *C*?

When tests yield profiles or other nonnumerical assessments and translation into letter grades is necessary, Stage 5 activities also help the teacher determine the correspondence between the profiles and letter grades.

Many teachers postpone setting a performance standard for a test until after they see how their students do. However, if one waits until after tests are given to plan how to interpret students' performance, the purpose for giving the test might be subverted. Using the labeling section of Macdonald's test to illustrate (Appendix), the teacher might decide that students must answer all four items correctly to demonstrate an acceptable level of understanding of the main idea of the passage. That is, the students should be able to perform this task perfectly if they are to be considered able to read and understand the main idea of the class texts. If the teacher finds that not one of her students answers all four correctly, it may be that none is ready for the sheltered science class. Lowering performance standards after giving the test would not change the science teacher's expectations for the students, but rather give the false impression that the students have the ability to understand the main idea of science texts.

Figure 5.

A Diagram of Stage 5 Procedures
Proposing Performance Standards

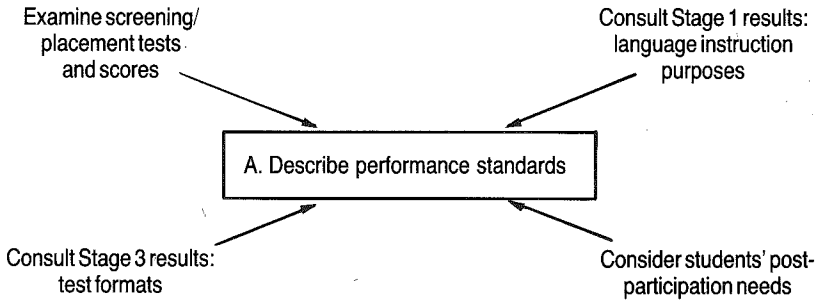
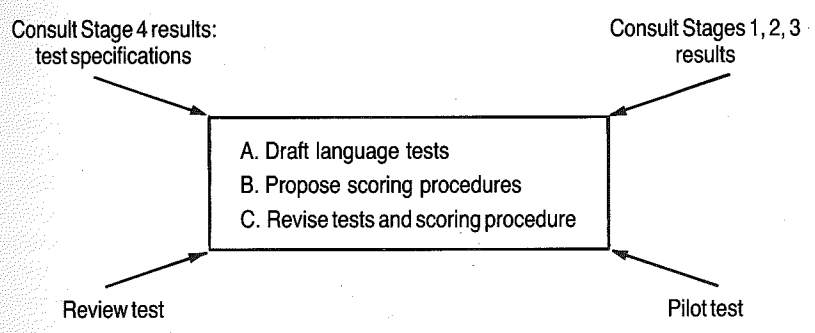


Figure 5 displays the procedures that should be used to determine performance standards. Deciding what performance characteristics or numerical score indicates an acceptable performance depends on not only the instructional goals of a particular language class, but also what students are expected to be able to do in subsequent language classes or language-use situations. This is true from the entry-level perspective as well. When defining what students should be able to do as they progress through a course, it is important to have a clear understanding of what they can do as they enter it. Consequently, the important steps in defining the performance standard for a test include determination of students' skills as they enter the class (usually through examining students' scores on whatever screening or placement test is used) and consideration of the language instruction goals and of students' language needs in situations for which the language class prepares them. Of course, the performance standard should reflect the format that was determined in earlier stages to be most appropriate for the particular CBI context in which the tests will be used.

Figure 6.

A Diagram of Stage 6 Procedures
Writing the Tests



The sixth stage of the guidelines (Figure 6) includes not only writing the test but also revising it. The results of either pilot testing or a critical review can guide revision. Pilot testing is the best way to collect information for revising a test. This procedure provides evidence to determine if the test instructions and tasks are clear enough, if the administration time is adequate, and if students actually interpret the items and tasks as the test writers intended. On objectively scored tests, item statistics such as *item difficulty* and *item discrimination* can easily be calculated. Finding a suitable group of may be difficult, but the information the process supplies makes the effort worthwhile.

Sometimes, however, pilot testing is simply not feasible. In these situations, the teacher/test writer should conduct especially thorough preadministration test review and revision. Ideally, a test should be

reviewed by a language teacher (other than the test writer) who is familiar with the particular situation for which the test is developed. Very often, colleagues are willing to exchange review responsibilities. When this is not possible, the test writer should review the test after allowing several days of objective distance to transpire. The reviewer should examine the items or tasks, making sure that they are appropriate and clear. The directions should be reviewed to be certain that they accurately delineate what the students have to do. Obviously, both pilot testing and test review require that the dark and dreary test-writing nights occur several days before the test administration date.

Stage 7, revising the tests, is performed after the tests are given and before scores are calculated or performance reports prepared. Despite the careful development procedures and the review process, there might be items or tasks which simply do not work—items or tasks that are confusing, ambiguous, or flawed in some other way. If problematic items or tasks are identified, they should be eliminated from the test. The results of those items or tasks should not contribute to students' scores or performance profiles. Although this means that the number of items or points might be changed from the original plan, it is only fair that students' test performance be assessed on the basis of good items rather than poor ones. Sometimes this results in an unexpected number of items—for example, a test that was intended to have 100 points might end up with 99 or 98. However, teachers who are troubled by a feeling of lost symmetry should be consoled by the fact that they have actually created more accurate measures of their students' abilities by eliminating poor items before calculating test scores.

Stage 8, the final stage of the guidelines, directs the teacher/test writer's efforts to determine the reliability and validity of the new test. An important consideration in this process is whether the test or test sections are objectively or subjectively scored. Objectively scored items are those which have only one correct response. In the matching section of Macdonald's test, for example, "dancing" can only be matched with "cerebellum," so one can say that this section is objectively scored. The essay, on the other hand, is subjectively scored. There is more than one correct answer—in fact, any individual's essay might be awarded the full 20 points even though each essay might be quite different. Both approaches to scoring are equally valuable although they are useful for different types of tasks.

Establishing the reliability of the scoring procedure is especially important for tests that are subjectively scored. One way to do this might be to ask the colleague who reviewed the test before it was given to score the tests as well. A correlation between the teacher/test writer's scores and the reviewer's scores establishes *interrater reliability,* an estimate of the consistency of scoring procedure across different scorers. Another way that consistency can be examined is to estimate *intrarater reliability.* To do this, the teacher/test writer scores the entire set of tests once, then scores them again perhaps the next day without consulting the first rating. While the teacher/test writer might not find perfect agreement between the first and second ratings (a correlation of 1.00), the scoring procedure should be clear enough to yield a high degree of consistency. Intrarater reliability lower than approximately .80 indicates that there is a serious problem with the consistency of the teacher/test writer's scores. The scoring procedure should therefore be modified to improve the consistency before reporting the students' scores.

Stage 8 also outlines steps to ensure the validity of a test; that is, whether it measures what it is intended to measure and measures it comprehensively. Expert review is one manner in which the validity of a test is estimated. The same reviewer who examined the test directions and content can be asked to make judgments regarding the appropriateness of the test content and the extent to which the test measures enough of whatever concept or skill it is designed to assess. For example, Macdonald indicates in her specifications that the test is intended to measure students' recognition of the main ideas in a reading passage. Does the first section of the test, the labeling task, require students to have understood the main idea of the passage (the name, position, and function of the four main parts of the central nervous system)? Not really, since the students do not need to understand the function of the parts to find and label them correctly. If this were the only task on the test, the test's validity would be weak. While the labeling task might require recognition of these four important parts and their location in the central nervous system, in terms of comprehensiveness, the test would fall short because it does not measure the students' understanding of the function of these parts. Inclusion of the second (matching) and fourth (essay) tasks increases the validity of the test with regard to its comprehensiveness. These tasks require the students to demonstrate their understanding of the function of the various parts of the brain as well as their location and labels.

Conclusion

Writing appropriate content-based language tests that are reliable and valid demands a commitment of time and care. The guidelines outlined in this article are not a shortcut to test writing—they do not produce instant tests. Teachers who follow the guidelines will devote long hours to creating their tests, just as they did before using the guidelines. However, they will be able to feel a greater sense of assurance in their tests' appropriateness, reliability, and validity as well as in the extent to which the tests measure their students' progress in both language and content mastery. ■

References

Macdonald, Elizabeth. (1991). *High school science testing project.* Project presented for Education 534: Language Testing, Monterey Institute of International Studies, Monterey, CA.

Popham, W. J. (1978). *Criterion-referenced measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Popham, W. J. (1981). *Modern educational measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Turner, J. L. (1991). *An adaptive model for the development of measures of language achievement in content-based language programs.* Unpublished doctoral dissertation, University of California, Los Angeles.

Appendix

Sample Content-Based Language Test

Instructions: Read the following passage carefully. As you read, you may want to make notes or circle important information that is in the text. When you have finished reading, you may begin the test. During the test, you should feel free to go back and reread the passage. Most of the information that you will need to answer the questions is in the passage itself.

The Central Nervous System

The central nervous system controls the human body. It's like the captain of a ship. Our brain is part of the central nervous system. It directs and controls everything that the human body does. There are four parts of the central nervous system: (a) the cerebrum, (b) the cerebellum, (c) the medulla, and (d) the spinal cord. The cerebrum, the cerebellum, and the medulla are all located in the brain. The spinal cord goes from the base of the brain down one's back. All of the different parts of the central nervous system have different functions.

The cerebrum is the largest part of the brain. It is the part of the brain that controls the senses, that is, seeing, hearing, feeling, tasting, and touching. It controls thinking and memory. People with good memories can remember many things. It also controls voluntary movement. Voluntary movement is movement that you choose to make. It is movement that you can control. Walking and talking are examples of voluntary movement.

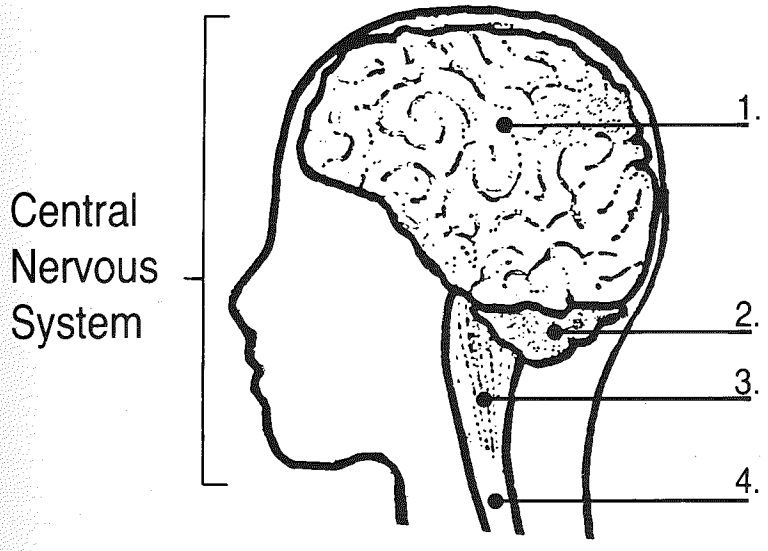
The cerebellum is located at the base of the cerebrum. It controls our sense of balance. If we didn't have balance, we would fall down. The cerebellum also controls coordination. Coordination is the ability to have all the different parts of one's body move and work together. Dancers and athletes, for example, must have good coordination.

The medulla controls involuntary movement. It is found in between the cerebellum and the spinal cord. It controls things that your body does without thinking. For example, it controls how you breathe, how your heart beats, and when you blink your eyes.

The spinal cord is the part of the central nervous system which carries information and messages to and from the brain. The spinal cord goes from the base of the neck, down the back. It is like a telephone wire. The messages and information that it carries are called impulses. These impulses must go through the spinal cord in order to get to the brain. The brain is able to send messages back to the body. These messages from the brain also must go through the spinal cord. If messages cannot go through the spinal cord, then the person is paralyzed. Often people who are paralyzed cannot move or talk.

Instructions:

1. Using the diagram below, label the four major parts of the central nervous system.



2. Draw a line from each part of the central nervous system to the activities that it controls.

Central Nervous System Part	Activity
Example:	Walking
	a. Talking
Cerebrum	b. Feeling cold
	c. Breathing
Cerebellum	d. Dancing
	e. Solving a math problem
Medulla	f. Sweating
	g. Telling a story
Spinal cord	h. Carrying impulses

3. Complete the following sentences using the most correct vocabulary word from the list.

voluntary	to control	memory
coordination	to be located	involuntary
paralyzed		

a. In order to play sports, you need good _____.
b. The medulla _____ in between the cerebellum and the spinal cord.
c. Movements that you control are _____.
d. Coughing is an example of _____ movement.
e. People whose spinal cords are damaged are often _____.
f. A student who has a good _____ usually gets good grades.
g. Messages from the brain are carried through the spinal cord and _____ the body's activities.

4. Write a one-paragraph essay explaining what parts of the brain are most important when you are playing a sport. You may choose any sport—soccer, tennis, swimming, basketball, football, and so forth.

Note. From *High School Science Testing Project* by Elizabeth Macdonald, 1991. Project presented for Education 534: Language Testing, Monterey Institute of International Studies, Monterey, CA. Reprinted by permission.