7

Constructed-Response Tests

In this chapter, you're going to learn about constructed-response tests. To be really truthful, you're only going to learn about two kinds of *paper-and-pencil* constructed-response items—namely, short-answer items and essay items (including students' written compositions). Although I suspect you already know that "you can't tell a book by its cover," now you've discovered that a chapter's title doesn't always describe its contents accurately. Fortunately, as far as I know, there are no state or federal "truth-in-chapter-entitling" laws.

You might be wondering why your ordinarily honest, never-fib author has descended to this act of blatant mislabeling. Actually, it's just to keep your reading chores more manageable. In an earlier chapter, you learned that student-constructed responses can be obtained from a wide variety of item types. In this chapter, we'll be looking at two rather traditional forms of constructed-response items, both of them paper-and-pencil in nature. In the next chapter, the focus will be on *performance tests*, such as those that arise when we ask students to make oral presentations or supply comprehensive demonstrations of skills in class. After that, in Chapter 9, we'll be dealing with portfolio assessment and how portfolios are used for assessment purposes.

Actually, all three chapters could be lumped under the single description of *performance assessment* or *constructed-response measurement* because any time you assess your students by asking them to respond in other than a selected-response manner, the students are *constructing*; that is, they are *performing*. It's just that if I'd loaded all of that performance assessment stuff in a single chapter, you'd have thought you were experiencing a month-long TV mini-series. The self-check exercises would have been as long as a short novel, and you'd have been pondering the pondertime questions for years. Yes, it was my inherent kindness and concern for readers that led me to disregard accuracy when entitling this chapter.

The major payoff of all constructed-response items is that they elicit student responses more closely approximating the kinds of behavior students must dis-

play in real life. After students leave school, for example, their demands of daily living almost never require them to choose responses from four nicely arranged alternatives. And when was the last time, in normal conversation, that you were obliged to render a flock of true-false judgments about a set of statements? Yet, you may well be asked to make a brief oral presentation to your fellow teachers or to a parent group, or you may be asked to write a brief report for the school newspaper about your students' fieldtrip to city hall. Constructed-response tasks unquestionably coincide more closely with nonacademic tasks than do selected-response tasks.

As a practical matter, if the nature of a selected-response task is sufficiently close to what might be garnered from a constructed-response item, then you may wish to consider a selected-response assessment tactic to be a reasonable surrogate for a constructed-response assessment tactic. Selected-response tests are clearly much more efficient to score. And, because almost all teachers I know are busy folks, time-saving procedures are not to be scoffed at. Yet, there will be situations when you'll want to make inferences about your students' status when selected-response tests just won't fill the bill. For instance, if you wish to know what kind of a cursive writer Jamal is, then you'll have to let Jamal write cursively. A true-false test about *i* dotting and *t* crossing just doesn't cut it.

SHORT-ANSWER ITEMS

The first kind of constructed-response item we'll look at is the *short-answer item*. Short-answer items call for students to supply a word, a phrase, or a sentence in response to either a direct question or an incomplete statement. If an item asks students to come up with a fairly lengthy response, it would be considered an essay item, not a short-answer item. If the item asks students to supply only a single word, then it's a *really* short-answer item.

Short-answer items are suitable for assessing relatively simple kinds of learning outcomes such as those focused on students' acquisition of knowledge. If crafted carefully, however, short-answer items can measure substantially more challenging kinds of learning outcomes. The major advantage of short-answer items is that students need to *produce* a correct answer, not merely recognize it from a set of selected-response options. The level of partial knowledge that might allow a student to respond correctly to a choose-the-best-response item won't be sufficient when the student is required to produce a correct answer to a short-answer item.

The major drawback with short-answer items, as is true with all constructed-response items, is that students' responses are difficult to score. The longer the responses sought, the tougher it is to score them accurately. And inaccurate scoring, as we saw in Chapter 2, leads to reduced reliability, which, in turn, reduces the validity of the test-based inferences we make about students, which, in turn, reduces the quality of the decisions we base on those inferences. Educational measurement is much like the rest of life—it's simply loaded with trade-

1. Usually employ direct questions rather than incomplete statements, particularly for young students.
2. Structure the item so that a response should be concise.
3. Place blanks in the margin for direct questions or near the end of incomplete statements.
4. For incomplete statements, use only one or, at most, two blanks.
5. Make sure blanks for all items are equal in length.

FIGURE 7–1 Item-Writing Guidelines for Short-Answer Items

offs. When classroom teachers choose constructed-response tests, they must be willing to trade some scoring accuracy (that comes with selected-response approaches) for greater congruence between constructed-response assessment strategies and the kinds of student behaviors about which inferences are to be made.

In Figure 7–1 you will find five straightforward item-writing guidelines for short-answer items. Please look them over, then I'll briefly amplify each guideline and describe how it works.

Using Direct Questions Rather than Incomplete Statements

For young children, the direct question is a far more familiar format than the incomplete statement. Accordingly, such students will be less confused if direct questions are employed. Another reason that short-answer items should employ a direct-question format is that the use of direct questions typically forces the item writer to phrase the item so less ambiguity is present. With incomplete-statements formats, there's often too much temptation simply to delete words or phrases from statements the teacher finds in textbooks. To make sure that there isn't more than one correct answer to a short-answer item, it is often helpful if the item writer first decides on the correct answer, then builds a question or incomplete statement designed to elicit a *unique* correct response from knowledgeable students.

Nurturing Concise Responses

Responses to short-answer items, as might be inferred from what they're called, should be *short*. Thus, no matter whether you're eliciting responses that are words, symbols, phrases, or numbers, try to structure the item so that a brief response is clearly sought. Suppose you conjured up an incomplete statement item such as this: "An animal that walks on two feet is _____." There are all sorts of answers that a student might legitimately make to such an item. Moreover, some of those responses could be fairly lengthy. Now note how a slight restructuring of the item constrains the student: "An animal that walks on two feet is technically classified as _____." By the addition of the phrase

"technically classified as," the item writer has restricted the appropriate responses to only one—namely, "biped." If your short-answer items are trying to elicit students' phrases or sentences, you may wish to place word limits on each, or at least indicate in the test's directions that only a *short* one-sentence response is allowable for each item.

Always try to put yourself, mentally, inside the heads of your students and try to anticipate how they are apt to interpret an item. What this second guideline suggests is that you massage an item until it truly lives up to its name—that is, until it becomes a *bona fide* short-answer item.

Positioning Blanks

If you're using direct questions in your short-answer items, place the students' response areas for all items near the right-hand margin of the page, immediately after the item's questions. By doing so, you'll have all of a student's responses nicely lined up for scoring. If you're using incomplete statements, try to place the blank near the end of the statement, not near its beginning. A blank positioned too early in a sentence tends to confuse the students. For instance, notice how this too-early blank can lead to confusion: "The _____ is the governmental body which, based on the United States Constitution, must ratify all U.S. treaties with foreign nations." It would be better to use a direct question or to phrase the item as follows: "The governmental body which, based on the United States Constitution, must ratify all U.S. treaties with foreign nations is the _____."

Limiting Blanks

For incomplete-statement types of short-answer items, you should use only one or two blanks. Any more blanks and the item is labeled a "swiss-cheese item," or an item with holes galore. Here's a swiss-cheese item to illustrate how confusing a profusion of blanks can make what is otherwise a decent short-answer item: "After a series of major conflicts with natural disasters, in the year ____, the explorers _____ and _____, accompanied by their _____, discovered _____." The student who could supply correct answers to such a flawed short-answer item should also be regarded as a truly successful explorer.

Inducing Linear Equality

Too often in short-answer items, a beginning item writer will give away the game by varying the length of the answer blanks so that short lines are used when short answers are correct and long lines are used when lengthier answers are correct. This practice tosses unintended clues to students and so should be avoided. In the interest of linear egalitarianism, not to mention decent item writing, keep all blanks for short-answer items equal in length. Be sure, however, that the length of the answer spaces provided is sufficient for students' responses—in other words, not so skimpy that students have to cram their answers in an illegible fashion.

In review, short-answer items are the most simple form of constructed-response items, but they can help teachers measure important skills and knowledge. Because such items seek students' constructed rather than selected responses, they can be employed to tap some genuinely higher-order skills. Although students' responses to short-answer items are more difficult to score than are their answers to selected-response items, the scoring of such items isn't all that difficult. That's because short-answer items, by definition, should elicit only *short* answers.

ESSAY ITEMS: DEVELOPMENT

The essay item is surely the most commonly used form of constructed-response assessment item. Any time teachers ask their students to churn out a paragraph or two on what the students know about Topic X or to compose an original composition describing their "Favorite Day," an essay item is being used. Essay items are particularly useful in gauging a student's ability to synthesize, evaluate, and compose. Such items have a wide variety of applications in most teachers' classrooms.

A special form of the essay item is the *writing sample*—when teachers ask students to generate a written composition in an attempt to measure students' composition skills. Because the procedures employed to construct items for such writing samples and, thereafter, for scoring students' compositions are so similar to the procedures employed to create and score responses to any kind of essay item, we'll treat writing samples and other kinds of essay items all at one time in this chapter.

For assessing certain kinds of complex learning outcomes, the essay item is the hands-down winner. It clearly wins out when you're trying to see how well students can create original compositions. Yet, there are a fair number of drawbacks associated with essay items, and if you're going to consider using essay items in your own classroom, you have to know the weaknesses as well as the strengths of this item type.

One difficulty with essay items is that they're more difficult to write—at least write properly—than is generally thought. I must confess that as a first-year high school teacher I sometimes conjured up essay items while walking to school, then slapped them up in the chalkboard so that I created almost instant essay exams. At the time, I thought my essay items were pretty good. Such is the pride of youth and the product of ignorance. I'm glad that I have no record of those items. In retrospect, I assume they were pretty putrid. I now know that generating a really good essay item is a tough task, a task that could not be accomplished in an instant. You'll see that from the item-writing rules to be presented shortly. It takes time to create a solid essay item. You'll need to find time to create suitable essay items for your own classroom assessments.

The most serious problem with essay items, however, is the difficulty that teachers have in scoring students' responses reliably. Let's say you use a six-item

Forests or Trees?

Allison Allen is a brand-new English teacher assigned to work with seventh-grade and eighth-grade students at Dubois Junior High School. Allison has taken part in a state-sponsored summer workshop that emphasizes "writing as a process." Coupled with what she learned while completing her teacher education program, Allison is confident that she can effectively employ techniques such as brainstorming, outlining, early drafts, peer critiquing, and multiple revisions. She assumes that her students will acquire not only competence in their composition capabilities but also confidence about their possession of those capabilities. What Allison's preparation failed to address, however, was how to grade her students' compositions.

Two experienced English teachers at Dubois Junior High have gone out of their way to help Allison get through her first year as a teacher. Mrs. Miller and Ms. Stovall have both been quite helpful during the early weeks of the school year. However, when Allison asked them one day during lunch how she should judge the quality of her students' compositions, two decisively different messages were given.

Ms. Stovall strongly endorsed *holistic* grading of compositions—that is, a general appraisal of each composition as a whole. Although Ms. Stovall bases her holistic grading scheme on a set of explicit criteria, she believes a single "gestalt" grade should be given so that "one's vision of the forest is not obscured by tree-counting."

Arguing with equal vigor, Mrs. Miller urged Allison to adopt *analytic* appraisals of her students' compositions. "By supplying your students with a criterion-by-criterion judgment of their work," she contended, "each student will be able to know precisely what's good and what isn't." (It was evident during the fairly heated interchanges that Mrs. Miller and Ms. Stovall had disagreed about this topic in the past.) Mrs. Miller concluded her remarks by saying, "Forget about that forest-and-trees metaphor, Allison. What we're talking about here is clarity!"

If you were Allison, how would you decide to judge the quality of your students' compositions?

essay test to measure your students' ability to solve certain kinds of problems in social studies. Suppose that, by some stroke of measurement magic, all your students' responses could be transformed into typed manuscript form so you couldn't tell which response came from which student. Let's say that you were asked to score the complete set of responses twice. What do you think is the likelihood your two sets of scores would be consistent? Well, experience suggests that most teachers aren't able to produce very reliable results when they score students' essay responses. The task in this instance, of course, is to *increase* the reliability of your scoring efforts so that you're not distorting the validity of the score-based inferences you want to make on the basis of your students' responses.

Creating Essay Items

Because the scoring of essay responses (and students' compositions) is such an important topic, you'll soon be getting a separate set of guidelines on how to score responses to such items. The more complex the nature of students' con-

1. Convey to students a clear idea regarding the extensiveness of the response desired.
2. Construct items so that the student's task is explicitly described.
3. Provide students with the approximate time to be expended on each item as well as each item's value.
4. Do not employ optional items.
5. Precursively judge an item's quality by composing, mentally or in writing, a possible response.

FIGURE 7–2 Item-Writing Guidelines for Essay Items

structed responses become, as you'll see in the next two chapters, the more attention you'll need to lavish on scoring. You can't score responses to items that you haven't written, however, so let's look in on Figure 7–2 where five guidelines for the construction of essay items are listed.

Communicating the Desired Extensiveness of Students' Responses

It is sometimes thought that if teachers decide to use essay items, students have total freedom of response. On the contrary, teachers can structure essay items so that students produce (1) barely more than they would for a short-answer item or (2) extremely lengthy responses. The two types of essay items reflecting this distinction in the desired extensiveness of student's responses are described as restricted-response items and extended-response items.

Restricted-response items decisively limit the form and content of students' responses. For example, a restricted-response item in a health education class might ask students the following: "Describe the three most common ways that HIV, the AIDS virus, is transmitted. Take no more than 25 words to describe each method of transmission." In this example, the number of HIV transmission methods was specified, as was the maximum length for each transmission method's description.

In contrast, an *extended-response item* provides students with far more latitude in responding. Here's an example of an extended-response item from a social studies class: "Identify the chief factors contributing to the enormous United States financial deficit of the 1980s and 1990s. Having identified those factors, decide which factors, if any, have been explicitly addressed by the U.S legislature and/or executive branches in the last five years. Finally, critically evaluate the likelihood that any currently proposed remedies will bring about significant reductions in the U.S. national debt." A decent response to such an extended-response item not only should get high marks from the teacher but might also be the springboard for a successful career in politics.

One technique that teachers commonly use to limit students' responses is to provide a certain amount of space on the test paper or in their students' response booklets. For instance, the teacher might direct students to "Use no more than two sheets (both sides) in your blue books to respond to each test item." Although the space-limiting ploy is an easy one to implement, it really disadvantages students

who write in a large-letter, scrawling fashion. Whereas such large-letter students may only be able to cram a few paragraphs onto a page, those students who write in a small, scrunched-up style may be able to produce a short novella in the same space.

This first guideline asks you to think carefully about whether the inference at issue that you wish to make about your students is best serviced by students' responses to (1) more essay items requiring shorter responses or (2) fewer essay items requiring extensive responses. Having made that decision, then be sure to convey to your students a clear picture of the degree of extensiveness you're looking for in their responses.

Describing Students' Tasks

Students will find it difficult to construct responses to tasks if they don't understand what the tasks are. Moreover, students' responses to badly understood tasks are almost certain to yield flawed inferences by teachers. The most important part of an essay item is, without question, the description of the *assessment task*. It is the assessment task that students respond to when they generate essays. Clearly, then, poorly described assessment tasks will yield many off-target responses that, had the student truly understood what was being sought, might have been more appropriate.

There are numerous labels used to represent the assessment task in an essay item. Sometimes it's simply called the *task*, the *charge*, or perhaps the *assignment*. In essay items that are aimed at eliciting student compositions, the assessment task is often referred to as the *prompt*. No matter how the assessment task is labeled, if you're a teacher who is using essay items, you must make sure that the nature of the task is really set forth clearly for your students. Put yourself, at least hypothetically, in the student's seat and see if, with the level of knowledge possessed by most of your students, the nature of the assessment task is really apt to be understood.

To illustrate, if you wrote the following essay item, there's little doubt that your students' assessment task would have been badly described: "In 500 words or less, discuss democracy in Latin America." In contrast, notice in the following item how much more clearly the assessment task is set forth: "Describe how the checks and balances provisions in the U.S. Constitution were believed by the Constitution's framers to be a powerful means to preserve democracy (300–500 words)."

Providing Time-Limit and Item-Value Guidance

When teachers create an examination consisting of essay items, they often have an idea regarding which items will take more of the students' time. But students don't know what's in the teacher's head. As a consequence, some students will lavish loads of attention on items that the teacher thought warranted only modest effort, yet devote little time to items that the teacher thought deserved substantial

attention. Similarly, sometimes teachers will want to weight certain items more heavily than others. Again, if students are unaware of which items count most, they may toss loads of rhetoric at the low-value items and have insufficient time to give more than a trifling response to the high-value items.

To avoid these problems, there's quite a straightforward solution—namely, letting students in on the secret. If there are any differences among items in point value or in the time students should spend on them, simply provide this information in the directions or, perhaps parenthetically, at the end of each item. Students will appreciate such clarifications of your expectations.

Avoiding Optionality

It's fairly common practice among teachers who use essay examinations to provide students with a certain number of items, then let each student choose to answer fewer than the number of items presented. For example, the teacher might allow students to "choose any five of the seven essay items presented." Students, of course, really groove on such an assessment procedure because they can respond to items for which they're well prepared and avoid those items for which they're inadequately prepared. Yet, other than inducing student glee, this optional-items classroom assessment scheme has little going for it.

When students select different items from a menu of possible items, they are actually responding to different examinations. As a consequence, it is impossible to judge their performances on some kind of common scale. Remember, as a classroom teacher you'll be trying to make better educational decisions about your students by relying on test-based inferences regarding those students. It's tough enough to make a decent test-based inference when you have only one test to consider. It's infinitely more difficult to make such inferences when you are faced with a medley of different tests because you allow your students to engage in a mix-and-match measurement procedure.

In most cases, teachers rely on an optional-items procedure with essay items when they're uncertain about the importance of the content measured by their examinations' items. Such uncertainty gives rise to the use of optional items because the teacher is not clearheaded about the inferences for which the examination's results will be used. If you spell out those inferences crisply, prior to the examination, you will usually find you'll have no need for optional item selection in your essay examinations.

Previewing Students' Responses

After you've constructed an essay item for one of your classroom assessments, there's a quick way to get a preliminary fix on whether the item is a winner or loser. Simply toss yourself, psychologically, into the head of one of your typical students, then anticipate how such a student would respond to the item. If you have time, and are inclined to do so, you could try writing a response that the student might produce to the item. More often than not, because you'll be too

busy to conjure up such fictitious responses in written form, you might try to compose a mental response to the item on behalf of the typical student you've selected. An early run-through of how a student might respond to an item can often help you identify deficits in the item because when you put yourself, even hypothetically, on the other side of the teacher's desk, you'll sometimes discover shortcomings in items that you otherwise wouldn't have identified. Too many times I've seen teachers give birth to a set of essay questions, send them into battle on examination day, and only then discover that one or more of the items suffers severe genetic deficits. Mental previewing of likely student responses can help you detect such flaws while there's still time for repairs.

In review, we've looked at five guidelines for creating essay items. If you'll remember that all of these charming little collections of item-specific recommendations should be adhered to *in addition to* the five general item-writing commandments set forth in Chapter 6 (Figure 6–1), you'll probably be able to come up with a pretty fair set of essay items. Then, perish the thought, you'll have to score your students' responses to those items. That's what we'll be looking at next.

ESSAY ITEMS: SCORING STUDENTS' RESPONSES

If you'll recall what's to come in future chapters (and, of course, recall is the *lowest* level of Bloom's cognitive taxonomy), you'll be looking at how to evaluate students' responses to performance assessments in Chapter 8 and how to judge students' portfolios in Chapter 9. In short, you'll be learning much more about how to evaluate your students' performances in constructed-response assessments. Thus, to spread out the load a bit, in this chapter we'll be looking only at how to score students' responses to essay items (including tests of students' composition skills). You'll find that many of the suggestions for scoring students' constructed responses that you will encounter in the following two chapters will also be applicable when you're trying to judge your students' essay responses. But just to keep matters simple, let's look now at recommendations for scoring students' responses to essay items. In Figure 7–3, you'll find a set of five such guidelines.

1. Score responses holistically and/or analytically.
2. Prepare a tentative scoring key in advance of judging students' responses.
3. Make decisions regarding the importance of the mechanics of writing prior to scoring.
4. Score all responses to one item before scoring responses to the next item.
5. Insofar as possible, evaluate responses anonymously.

FIGURE 7–3 Guidelines for Scoring Responses to Essay Items

Choosing an Analytic and/or Holistic Scoring Approach

During the past decade or two, the measurement of students' composition skills by having students generate actual writing samples has become widespread. As a consequence of all this attention to students' compositions, educators have become far more skilled regarding how to evaluate students' written compositions. Fortunately, classroom teachers can use many of the procedures that have been identified and refined while educators scored thousands of students' compositions during statewide assessment extravaganzas. (In Texas, over a million students' essays were scored annually. Even for Texas, pardner, that's a pile of essays.)

A fair number of the lessons learned about scoring students' writing samples apply quite nicely to the scoring of responses to any kind of essay item. One of the most important of the scoring insights picked up from those large-scale scoring of students' compositions is that almost any type of student-constructed response can be scored either *holistically* or *analytically*. That's why the initial guideline in Figure 7–3 suggests you make an early-on decision whether you're going to score your students' responses to essay items using a holistic approach, an analytic approach or, perhaps, using a combination of the two scoring approaches. Let's look at how each of these two scoring strategies works.

A *holistic* scoring strategy, as its name suggests, focuses on the essay response (or written composition) as a whole. At one extreme of scoring rigor, the teacher can, in a fairly unsystematic manner, supply a "general impression" overall grade to each student's response. Or, in a more systematic fashion, the teacher can isolate, in advance of scoring, those evaluative factors that should be attended to in order to arrive at a single, overall score per essay. Generally, a score range of four to six points is used for each response. (Some scoring schemes have a few more points, some a few less.) A teacher, then, after considering whatever factors should be attended to in a given item, will give a score to each student's response. In Figure 7–4, you will see a set of evaluative factors that teachers might use in holistically scoring a student's written composition. In Figure 7–5 are fair factors that a speech teacher might employ in holistically scoring a response to an essay item used in a debate class.

For scoring a composition intended to reflect students' writing prowess:

- Organization
- Communicative Clarity
- Adaptation to Audience
- Word Choice
- Mechanics (spelling, capitalization, punctuation)

FIGURE 7–4 Illustrative Evaluative Factors That Could Be Considered When Scoring Students' Essay Responses Holistically

For scoring a response to an essay item dealing with rebuttal preparation:
- Anticipation of Opponent's Positive Points
- Support for One's Own Points Attacked by Opponents
- Isolation of Suitably Compelling Examples
- Preparation of a "Spontaneous" Conclusion

FIGURE 7–5 Potential Evaluative Factors That Could Be Used When Scoring Students' Essay Responses in a Debate Class

When teachers score students' responses holistically, they do *not* dole out points-per-factor for a student's response. Rather, the teacher keeps in mind evaluative factors such as those set forth in Figures 7–4 and 7–5. The speech teacher, for instance, while looking at the student's essay response to a question regarding how someone should engage in effective rebuttal preparation, will not necessarily penalize a student who overlooks one of the four evaluative factors cited in Figure 7–5. The response as a whole may lack one point, yet otherwise represent a really wonderful response. Evaluative factors such as those illustrated in Figures 7–4 and 7–5 simply dance around in the teacher's head when the teacher scores students' essay responses holistically.

In contrast, an *analytic* scoring scheme strives to be a fine-grained, specific point-allocation approach. Suppose, for example, that instead of using the holistic method of scoring students' compositions represented in Figures 7–4 and 7–5, a teacher chose to employ an analytic method of scoring students' compositions. Under those circumstances, a scoring guide such as the example in Figure 7–6 might be used by the teacher. Note that, for each factor in the guide, the teacher must award 0, 1, or 2 points. The lowest overall score for a student's composition, therefore, would be 0, whereas the highest overall score from a student's composition would be 10 (that is, 2 points times 5 factors).

The advantage of an analytic scoring system is that it can help you identify the specific strengths and weaknesses of your students' performances and there-

Factor	Unacceptable (0 points)	Satisfactory (1 point)	Outstanding (2 points)
1. Organization	___	✓	___
2. Communicative Clarity	___	___	✓
3. Audience Adaptation	___	✓	___
4. Word Choice	___	___	✓
5. Mechanics	___	✓	___

Total Score = 7

FIGURE 7–6 An Illustrative Guide for Analytically Scoring a Student's Written Composition

fore communicate such diagnoses to students in a pinpointed fashion. The downside of analytic scoring is that a teacher sometimes becomes so attentive to the subpoints in a scoring system that, almost literally, the forest (overall quality) can't be seen because of a focus on individual trees (the separate scoring criteria). In less metaphoric language, the teacher will miss the communication of the student's response "as a whole" because of excessive attention to a host of individual evaluative factors.

One middle-of-the-road scoring approach can be seen when teachers initially grade all students' responses holistically, then return for an analytic scoring only of those responses that were judged, overall, to be unsatisfactory. After the analytic scoring of the unsatisfactory responses, the teacher then relays more fine-grained diagnostic information to those students whose unsatisfactory responses were analytically scored.

This initial guideline for scoring students' essay responses applies to the scoring of responses to all kinds of essay items. As always, your decision about whether to opt for holistic or analytic scoring should flow directly from the use to which you'll be putting the test results. Putting it another way, your choice of scoring approach will depend on the educational decision linked to the test's results.

Devising a Tentative Scoring Key

No matter what sort of approach you opt for in scoring your students' essay responses, you'll find that it will be useful to develop a tentative scoring key for responses to each item *in advance* of actually scoring students' responses. Such tentative scoring schemes are almost certain to be revised on the basis of your scoring of actual student papers, but that's to be anticipated. If you wait until you commence scoring your students' essay responses, there's too much likelihood that you'll be unduly influenced by the responses of the first few students whose papers you grade. If those papers are atypical, the resultant scoring scheme is apt to be unsound. It is far better to think through, at least tentatively, what you really hope students will supply in their responses, then modify the scoring key if unanticipated responses from students suggest that alterations are requisite.

If you don't have a tentative scoring key in place, there's a great likelihood that you'll be influenced by such factors as a student's vocabulary or writing style even though, in reality, such variables may be of little importance to you. Advance exploration of the evaluative criteria you intend to employ, either holistically or analytically, is a winning idea when scoring responses to essay items.

Deciding Early about the Importance of Mechanics

Few things influence scorers of students' essay responses as much as the mechanics of writing employed in the response. If the student displays subpar spelling, chaotic capitalization, and poor punctuation, it's pretty tough for a scorer of the student's response not to be influenced adversely. In some instances, of course,

mechanics of writing certainly do play a meaningful role in scoring students' performances. For instance, suppose you're scoring students' written responses to a task of writing an application letter for a position on a local newspaper. In such an instance, it is clear that mechanics of writing would be pretty important when judging the student's response. But in a chemistry class, perhaps the teacher cares less about such factors when scoring students' essay responses to a problem-solving task. The third guideline simply suggests that you make up your mind about this issue early in the process so that, if mechanics aren't all that important to you, you don't let your students' writing mechanics subconsciously influence the way you score their responses.

Scoring One Item at a Time

If you're using an essay examination with more than one item, be sure to score all of your students' responses to one item, then score all of their responses to the next item, and so on. Do *not* score all responses of a given student, then go on to the next student's paper. There's too much danger that a student's responses to early items will unduly influence your scoring of responses to subsequent items. If you score all responses to item number 1, then move on to the responses to item number 2, you can eliminate this tendency. In addition, the scoring will actually go a bit quicker because you won't need to shift evaluative criteria between items. Adhering to this fourth guideline will invariably lead to more consistent scoring, hence to more accurate response-based inferences about your students. There'll be more paper shuffling than you might prefer, but the increased accuracy of your scoring will be worth it. (Besides, you'll be getting a smidge of psychomotor exercise.)

Striving for Anonymity

Because I've been a teacher, I know all too well how quickly teachers can identify their students' writing styles, particularly those students who have especially distinctive styles such as the "scrawlers," the "petite letter-size crew," and those who dot their *i*s with half-moons or cross their *t*s with lightning bolts. Yet, insofar as you can, try not to know whose responses you're scoring. One simple way to help in that effort is to ask students to write their names on the reverse side of the last sheet of the examination in the response booklet. Try not to peek at the students' names until you've scored all of the exams.

I used such an approach for three decades of scoring graduate students' essay examinations at UCLA. It worked fairly well. Sometimes I was really surprised because students who had appeared to be knowledgeable during class discussions displayed just the opposite on the exams, while several Silent Sarahs and Quiet Quentins came up with really solid exam performances. I'm sure that if I had known whose papers I had been grading, I would have been improperly influenced by my classroom-based perceptions of different students' abilities. I am not suggesting that you shouldn't use students' classroom discussions as part

of your evaluation system. Rather, I'm advising you that classroom-based perceptions of students can sometimes cloud your scoring of essay responses. That's one strong reason for you to strive for anonymous grading.

In review, we've considered five guidelines for scoring students' responses to essay examinations. If your classroom assessment procedures involve any essay items, you'll find that these five practical guidelines will go a long way in helping you come up with consistent scores for your students' responses. And consistency, as you learned in Chapter 2, is something that makes psychometricians mildly euphoric.

Now, after you've completed this chapter's usual end-of-chapter stuff, in the next two chapters you'll learn about two less common forms of constructed-response items. You'll learn how to create and score performance assessments and to use students' portfolios in classroom assessments. You'll find that what we've been dealing with in this chapter will serve as a useful springboard to the content of the next two chapters.

WHAT DO CLASSROOM TEACHERS REALLY NEED TO KNOW ABOUT CONSTRUCTED-RESPONSE TESTS?

At the close of a chapter dealing largely with the nuts and bolts of creating and scoring written constructed-response tests, you probably expect to be told that you really need to internalize all those nifty little guidelines so that, when you spin out your own short-answer and essay items, you'll elicit student responses that you can score accurately. Well, that's not a terrible aspiration, but there's really a more important insight you need to walk away with after reading the chapter. That insight, not surprisingly, derives from the central purpose of classroom assessment—namely, to draw accurate inferences about students' status so that you can make more appropriate educational decisions. What you really need to know about short-answer items and essay items is that you should use them as part of your classroom assessment procedures if you want to make inferences about your students that those students' responses to such items would support.

Putting it another way, if you, as a classroom teacher, want to determine if your students have the skill and/or knowledge that can be best measured by short-answer or essay items, then you need to refresh your memory regarding how to avoid serious item construction or response scoring errors. A review of the guidelines presented in Figures 7–1, 7–2, and 7–3 should give you the brushup that you need. Don't believe that you are obligated to use short-answer or essay items simply because you now know a bit more about how to crank them out. If you're interested in the extensiveness of your students' *knowledge* regarding Topic Z, it may be far more efficient to employ fairly low-level selected-response kinds of items. If, however, you really want to make inferences about your students' abilities to perform the kinds of tasks represented by short-answer and essay items, then the guidelines provided in the chapter should be consulted.

CHAPTER SUMMARY

After a fairly elaborate, guilt-induced apology for the chapter's mislabeling, the chapter started off with a description of short-answer items accompanied by a set of guidelines (Figure 7–1) regarding how to write short-answer items. The chapter then took up essay items and indicated that, although students' written compositions constitute a particular kind of essay response, most of the recommendations for constructing essay items and for scoring students' responses were the same, whether measuring students' composition skills or skills in subject areas other than language arts. Guidelines were provided for writing essay items (Figure 7–2) and for scoring students' responses to essay items (Figure 7–3). The chapter was concluded with the suggestion that much of the content to be treated in the following two chapters, because those chapters also focus on constructed-response assessment schemes, will relate to the creation and scoring of short-answer and essay items.

SELF-CHECK

This chapter's self-check exercises will start off with a brief set of illustrative short-answer and essay items. Your task is to identify whether any of the chapter's item-writing guidelines have been violated. If so, you are to indicate what the violated guideline was or, if two or more guidelines have been flaunted, which ones those were. If not, indicate that no guidelines have been violated. Because you'll be responding with brief constructed responses, what kind of test items do you suppose you're about to mess with?

Oh yes, if you want to, you might like to engage in a quick review of the item-writing guidelines in Figures 7–1 and 7–2. It's not necessary, but some people are compulsive when they think their self-worth is going to be under scrutiny.

1. (The following item was written for fifth-graders.) Before the United States became an independent nation, most of the men and women who had come to settle were from which European nation or nations?

   _____ _____ _____

   Any violation(s)?_____

   _____ _____

   _____ _____

2. (This item is for high school students.) You have just viewed a videotape containing three widely seen television commercials. What is the one classic propaganda technique present in all three commercials? _____

Any violation(s)?_____

_____

_____

3. (This item is for eighth-grade students.) If a friend has asked you to engage in the use of alcohol or _____, and you wish to decline, you should _____ and _____ without delay.

Any violation(s)?_____

_____

_____

4. (This item was written for eleventh-grade English students.) In the space provided below and on the attached sheet, please compose a brief 200-word editorial in favor of the school district's expanded after-school tutorial program. The intended audience for your position statement consists of those people who read the local newspaper's editorial page. Because you have the entire class period to prepare your response, please use the scratch paper provided for a first draft, then revise your editorial before copying it on these sheets. The editorial will contribute 40 percent to your six-week's Persuasive Writing grade.

Any violation(s)?_____

_____

_____

5. (This item was written for sixth-graders.) Thinking back over the mathematics lesson and homework assignments you had during the past 12 weeks, what conclusions can you draw? Take no more than one page for your response.

Any violation(s)?_____

_____

_____

6. For the chapter's final self-check exercise, please read the description of how a fictitious teacher went about scoring an essay examination in a high school class called U.S. Government Service. Then, evaluate how well the teacher carried out the scoring.

   "I was using a four-item essay examination on which my students were to spend 10 to 15 minutes per item. Before scoring students' papers, I worked up a tentative analytic scoring guide for each item. I identified two factors that *had* to be addressed in each response in order for it to receive full credit. If a student used a factor fairly well, the response earned 5 points; if the factor was well used, the response earned 10 points, In all, then, with two factors per item, and 0 to 10 points per factor, I could give a student 0 points up to 80 points. My scoring keys, which placed no emphasis on a student's use of mechanics, worked so well that I didn't have to revise them.

   "I scored each student's response to all four items at a single sitting, and I found that I could interpret students' responses more sensitively because, from the student's name on the first page of the response booklet, I could tell whose responses I was grading. The entire scoring operation for the 29 students in my class took somewhat less than 90 minutes."

   What is your evaluation of this teacher's scoring procedures?

PONDERTIME

1. What do you think has been the *instructional* impact, if any, of the increasingly wide incorporation of student writing samples in the high-stakes educational achievement tests used in numerous states?
2. How would you contrast short-answer items and essay items with respect to the levels of cognitive behavior set forth in Bloom's taxonomy? Are there differences in the kinds of cognitive behavior elicited by the two item types? If so, what are they?
3. What do you see as the major weaknesses of short-answer and essay items? What are their major strengths?
4. What would influence you in deciding whether to use selected-response items or the kinds of constructed-response items treated in this chapter?
5. What kinds of student outcomes could not be assessed properly with short-answer items but could be assessed properly with essay items?

SELF-CHECK KEY

1. This item really doesn't violate any of the guidelines listed on Figure 7–1, but it's a weak item anyway because it violated one of the general-purpose item-writing commandments cited in Chapter 6. The item presents an ambiguous task to students insofar as it asks for a nation (or nations) from which "most" of the settlers came, then provides three blank spaces, thereby indicating that three, not one, response is sought even though, by definition, there is only one "most." An astute answer from you for this self-check item would have been something along these lines: "No violations, but cruddy nonetheless."
2. This item violates no guidelines and isn't really all that shabby.
3. This item violates a pair of guidelines. It uses too many blanks and the blanks are unequal in length. It would be difficult for students, given the nearly total ambiguity introduced by the initial blank, to know how to respond to that item. It is, in short, a loser item.
4. This item violates none of the chapter's item-writing guidelines for essay items. As you can note, the first three guidelines in Figure 7–2 have all been specifically followed in the item. It's a fairly decent prompt for a persuasive writing task.
5. This item violates several of the chapter's guidelines for writing essay items. Although it limits the student's response to one page, there is substantial ambiguity about the nature of the task that the teacher wants the students to undertake. Twelve weeks worth of mathematics can add up (a mathematical operation) to a pretty hefty pile of math. The number of conclusions that the student can legitimately draw are myriad. Thus, Guideline 2 is clearly violated because the student doesn't know what to do. Guideline 3 is also violated because there's no indication given regarding time limit or point value for the item. This item needs a good deal of refurbishing.
6. This fictitious teacher adhered to one of the guidelines in Figure 7–3 when a tentative analytic scoring key was used that dealt with the importance of a student's written mechanics. There were two serious guideline violations, however, in that the teacher didn't score items one at a time and also failed to score students' responses anonymously. On balance, therefore, you should have given a "thumbs-down" to this teacher's essay-scoring effort. The two serious violations are too likely to threaten the validity of any score-based inferences the teacher might wish to draw.

ADDITIONAL STUFF

Gronlund, N. E., and R. L. Linn. *Measurement and Evaluation in Teaching* (6th ed.). New York: Macmillan, 1990.

Haladyna, Thomas M. "The Effectiveness of Several Multiple-Choice Formats." *Applied Measurement in Education* 5, no. 1 (1992): 73–88.

Hopkins, Charles D., and Richard L. Antes. *Classroom Testing Construction* (2nd ed.). Itasca, IL: F. E. Peacock, 1989.

Kaplan, R. M. "Scoring Natural Language Free-Response Items." *Proceedings of the 33rd Annual Conference of the Military Testing Association*, 1992, 514–518.

Mehrens, W. A., and I. J. Lehmann. *Measurement and Evaluation in Education and Psychology* (4th ed.). New York: Holt, Rinehart & Winston, 1991.

Roid, G. H., and T. M. Haladyna. *A Technology for Test-Item Writing*. New York: Academic, 1982.

NONPRINT STUFF

Northwest Regional Laboratory. *Writing Assessment: Issues and Answers* (Videotape #NREL-7A). Los Angeles: IOX Assessment Associates.

Northwest Regional Laboratory. *Writing Assessment: Training in Analytical Scoring* (Videotape #NREL-8A). Los Angeles: IOX Assessment Associates.