1

Why Do We Test?

Most of us who chose to become educators did so in order to help children learn the things they ought to learn. Typically, we started off by wanting to be teachers. Then, after we'd taught for a while, some of us decided to tackle other educational challenges, such as becoming school administrators. But the dominant motive for first selecting an educational career is almost always to help students learn. Let's face it, few people opt to become educators as part of a get-rich-quick financial strategy.

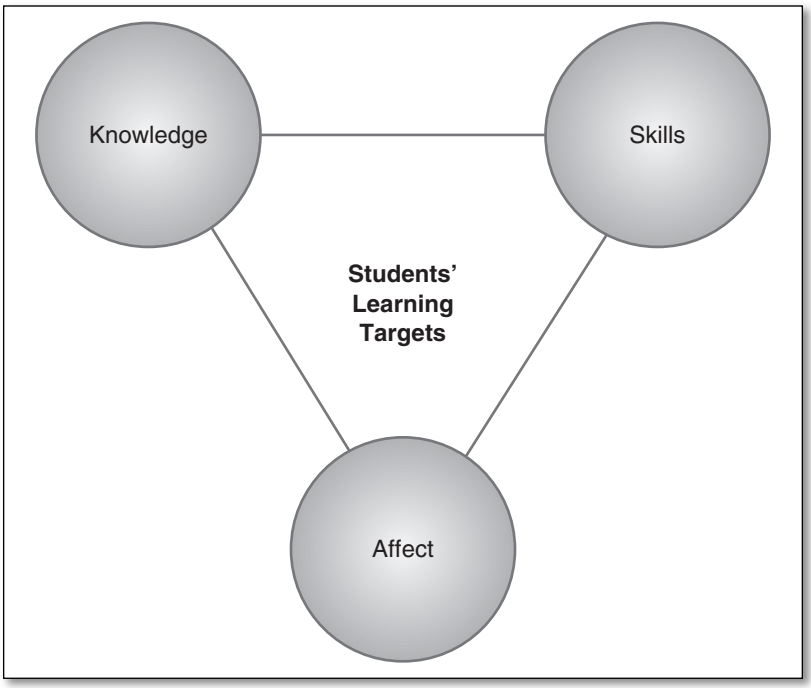
Okay, what is the nature of this "learning" we hope to promote in our students? Well, the things students ought to learn are, for the most part, skills and knowledge. The *skills* involved are usually intellectual skills, such as when students are able to compose a coherent essay. But children also need psychomotor skills, such as being able to use a computer's keyboard or discovering how to stay afloat while swimming. With respect to *knowledge,* there is a truly enormous collection of stuff that students need to know, for example, flocks of facts, tons of truths, and piles of principles. The more knowledgeable we

can make students, the more likely it is those students can then employ their knowledge to deal with the world around them. If we can help students become both skillful *and* knowledgeable, of course, this will supply pretty potent support for those students' future success—and for their personal happiness.

However, beyond knowledge and skills, we should also be attentive to students' *affect,* that is, we should be attentive to students' attitudes, interests, and values. Because a student's affective dispositions can have an enormous impact on that student's life, educators who fail to promote appropriate affect for their students are falling down on a significant educational responsibility. We definitely want our students to learn how to read with comprehension, but we should also want them to *enjoy* reading. We definitely want our students to master mathematics, but we should also want them *not to fear* mathematics. So, in addition to promoting students' attainment of suitable skills or their mastery of needed knowledge, we should be also nurture students' acquisition of defensible affective dispositions. (Later, in Chapter 9, we'll consider ways that educators can assess students' affect.)

In a very real sense, then, education is organized so as to promote students' achievement of the three kinds of outcomes you see represented in Figure 1.1, namely, students' becoming more knowledgeable, more skilled, and in possession of life-enhancing affective dispositions. Admittedly, much more attention is currently given to promoting students' skills and knowledge than to promoting students' affect, and this will probably always be so. However, it may be helpful for you to recognize that, whether you are currently a classroom teacher or a school administrator, the overriding goal of the educational system in which you function is to promote students' attainment of what's in the three circles set forth in Figure 1.1. I find it difficult to conceive of any educational setting in which teachers should not at least *consider* the possibility of influencing students' knowledge, skills, and affect.

Figure 1.1 The Three Outcomes Educators Seek for Their Students



The Role of Assessment

If the mission of educators is to get students to end up with appropriate knowledge, skills, and affect, then where does assessment come in? Indeed, a fundamental question that all educational leaders should be able to answer—both clearly and concisely—is, *Why do we test students?* The current chapter is intended to help you answer this question. But, first, let me try to clarify a few potential terminology tangles that might, if not addressed, impede us.

In the remainder of the book, I will interchangeably employ the terms *test*, *assessment*, and *measurement*. Most of the time, I'll be using *assessment* because it appears to be the term most currently favored by educators. This is probably because *measurement* is a term seen as at least a little off

putting, and when people use the word *test,* they often think of the sorts of traditional paper-and-pencil tests most of today's adults experienced when, as students ourselves, we went through school. Yet, especially in the last few years, we have been measuring students using considerably more diverse assessment procedures than represented by multiple-choice, true-false, and essay tests. Subtle definitional differences aside, however, when you understand the basic reason we "assess," "test," or "measure" students, you'll see that, regardless of which of these three labels is being employed, what's going on is pretty much the same thing.

As long as I've taken a short detour to do some terminology tightening, please allow me to clarify three more terms that will soon pop up in this chapter and will often be seen in later chapters. I'm referring to *curriculum, instruction,* and *evaluation.* You'll find that most educators understand what's meant by *instruction*—it's simply another label for *teaching.* And, *evaluation* is accurately thought by most educators to focus on how we determine the effectiveness of instruction, for instance, when teachers try to determine the quality of their own teaching. But there's a fair amount of confusion regarding the meaning of *curriculum.* Some educators think of curriculum as what is supposed to be learned by students; other educators regard curriculum as the materials used during instruction; and other educators think of curriculum as the actual activities that go on in class. Clearly, these definitions differ.

The way I'll be using the term *curriculum* in this book is as a label to describe educational *ends,* that is, the intended outcomes we want our students to achieve. Thus, the three types of learning targets seen in Figure 1.1, knowledge, skills, and affect, are all intended outcomes and, thus, can be thought of as *curricular aims,* that is, the skills, knowledge, and affect we aim for our students to attain. If curricular aims are the intended outcomes of education, then instruction (or, if you prefer, teaching) represents the *means* by which we intend to have students attain our chosen curricular ends. Along the way, we may wish to assess students during instruction to see if we need to make any changes in our current or immediately upcoming instructional

activities. And once instruction is over—that is, when a sequence of instructional activities has been concluded—there is the need to evaluate the success of the instruction so we can determine whether to modify this instructional sequence when we use it again with future students.

As you will see, it is because of the need to make curricular, instructional, and evaluative *decisions* that educators assess their students. That's right; we don't test for the sheer joy of testing or because "it is interesting." Instead, we assess students in order to make better *decisions* about the curricular ends we should be pursuing, the way our instruction is working, and—at the close of instruction—how successfully students have achieved our intended curricular aims. Education is a decision-making enterprise, and educational measurement helps educators make better decisions. In short, we test our students so we can make more appropriate decisions about how to educate them.

Coping With the Covert

But there's a complication. We can't, by looking at kids, tell what they know. We can't, by looking at kids, tell what their skills are. We can't, by looking at kids, tell what their affective dispositions are. The reason we are unable to visually discern students' knowledge, skills, and affect is that all three of those things are *hidden* from view, that is, all three are *covert.* No matter how long you might look at a student who's sitting, perhaps only three feet in front of you, and no matter how intently and carefully you scrutinize this student, you simply cannot tell what's going on inside the student's skull. The student's knowledge, skills, and affect are quite invisible.

As indicated earlier, the entire educational enterprise revolves around educators' making appropriate curricular, instructional, and evaluative decisions. Yet, how can these decisions be made defensibly if educators have no idea about their students' *current* knowledge, skills, and affect? And this, of course, is where educational assessment comes roaring to the rescue—for it is through the use of assessment that educators can arrive at reasonable conclusions about such unseen variables as

students' knowledge, skills, and affect. So, the answer to the Why do we test? question is that we employ test results to arrive at inferences—which we then use to make better decisions. Appropriate decisions enhance the quality of education we provide to students. And this is precisely the goal embraced by every school leader I've ever known, namely, to improve the quality of schooling provided to students.

For example, when students complete an information-focused quiz, then return it to the teacher, they are providing overt evidence about how much covert information they currently possess. Similarly, when students compose an original narrative essay as part of an annual statewide accountability test, they are also supplying overt evidence about their covert composition skills. And, when a classful of students complete an anonymous affective inventory indicating their current attitudes toward mathematics, they are also supplying overt evidence about the way they regard math.

When educators consider such overt evidence, they are then able to arrive at evidence-based *inferences* (or, if you prefer, they arrive at evidence-based *interpretations*) about educational variables that simply can't be seen. Assessment, in other words, permits educators to reach inferences about what's going on—unseen—inside the students who are being educated. Educational assessment, then, is fundamentally an inference-making enterprise. As you'll see in the next chapter, school leaders must understand that inferences are made by human beings and are not delivered ready-made, in cut-and-dried fashion by assessment instruments themselves. We'll dip into that important understanding when we look at what's called "assessment validity."

Curricular Decisions

But testing students, and figuring out what sorts of inferences must be made from test results, is not done without purpose. We do so to arrive at more suitable decisions regarding curriculum, instruction, and evaluation. To illustrate, let's look briefly at curricular decisions. It would be patently dull

witted for educators to teach students stuff those students already know. By using appropriate assessment techniques, we can figure out what skills, knowledge, and affect a group of students actually possess when they initially come to us—thereby allowing us to avoid the serious curricular sin of trying to teach students what they've already learned.

Another set of curricular decisions depends directly on whether students possess the necessary precursive skills and knowledge so that a teacher's pursuit of particular curricular aims makes instructional sense. Without assessment, we can't tell if kids possess the requisite precursors. But by using educational assessment procedures, teachers can arrive at reasonable inferences about whether their students have the needed precursors before deciding to tackle specific curricular aims. Curricular choices made in the absence of assessment-elicited evidence about students' current status are curricular choices almost certain to be flawed. Defensible curricular decision making depends on the availability of assessment-generated evidence about students' current status.

Instructional Decisions

Turning to instruction, there are loads of instructional decisions that can be made more sensibly by relying on evidence regarding students' current status, for example, the degree to which students are making progress in mastering a challenging cognitive skill. If a teacher is deciding whether to spend more instructional time on a particular curricular aim and, if so, how to spend it wisely, these are decisions that can clearly be made more wisely when the teacher has evidence regarding how well the students have already learned what they are supposed to be learning.

Later, in Chapter 8, we'll take a careful look at a particularly powerful use of classroom-assessment evidence to improve teachers' instructional decision making when we consider *formative assessment.* But what should be clear to you already is that there are all sorts of teachers' instructional decisions to be

made more astutely if only a teacher has access to assessment results enabling the teacher to arrive at sound inferences regarding students' current status.

Evaluative Decisions

Another prominent use of assessment evidence is seen when teachers set out to *evaluate* the success of this year's instructional procedures in order to decide whether to modify those procedures in the future when new students rumble into the classroom. Although we can regard this sort of assessment-abetted activity as exclusively evaluative rather than instructional, it should be evident that when teachers evaluate their teaching—with the prospect of fixing any flaws in it—this evaluative activity has all sorts of implications for subsequent instructional decisions.

Because, at bottom, we should evaluate the quality of instruction according to its payoff in changing students, there are numerous instances when those who evaluate instruction will have need for assessment evidence. You can't tell what changes have taken place inside students without collecting evidence about instruction's impact on a swarm of educationally relevant variables that we simply can't see.

The Multiple-Measures Myth

Any educator who has done a spot of serious thinking about assessment realizes there is peril in basing an inference about a student on only one test. Not surprisingly, therefore, during the last decade we have seen increasing numbers of educational policymakers calling for the use of "multiple measures" when making important decisions about individual students or groups of educators. It is thought, not unreasonably, that a solo indicator of quality will surely lead to defensible decisions. Indeed, the demand for multiple measures has become almost a measurement mantra for some educational organizations who rail against reliance on only one measuring device, especially when important outcomes are tied to students' test scores.

Let's remember, however, when we try to arrive at inferences about students' covert status, a single *good* measurement device is almost always better than four or five *bad* measurement devices. Simply adding additional evidence-gathering procedures does not ensure more appropriate inferences

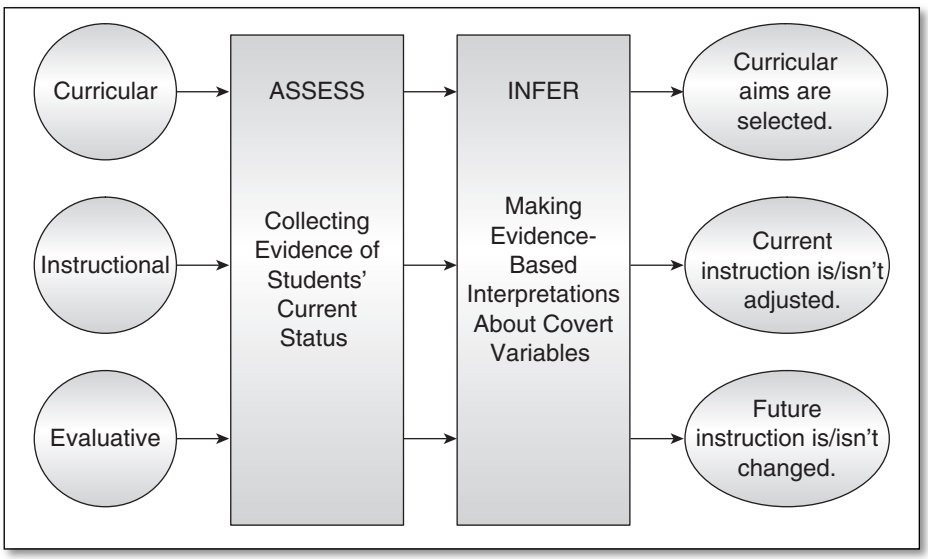
and decisions. Thus, when a school leader encounters a chorus of educators chanting for the use of multiple measures, the school leader should quickly pose the following question: What additional measures do you have in mind and, of course, how good are they?

Decisions, Decisions, Decisions

School leaders must understand, then, the only defensible reason for assessment of any sort is that it helps educators make better decisions about what to teach kids and how to teach them most effectively. Moreover, as you have seen, the evidence on whose basis these "better decisions" are made almost always requires the use of educational assessment.

Please consider Figure 1.2 for a moment. What you'll find depicted graphically is the way that the two mainstay operations of educational assessment—namely, assessing and inferring—are always the same, irrespective of whether the decision to be made is curricular, instructional, or evaluative. Let's look at what's going on in Figure 1.2.

Figure 1.2 Curricular, Instructional, and Evaluative Decision-Making Applications of Educational Assessment's Assess-Infer Essence



First off, please note the two center rectangles, identified as "Educational Assessment's Essence." As you can see in the rectangle at the left, this is when we actually assess students by using a potentially wide variety of measurement ploys such as formal paper-and-pencil tests or, perhaps, more innovative assessment procedures such as portfolio assessment. Then, in the right-hand rectangle, based on the overt evidence garnered from such measurement techniques, we arrive at inferences about the covert variables in which we are interested (such as students' knowledge of significant historical events or their algebraic skills). It is this *assess-infer* operation that we rely on to support a wide variety of educationally relevant decisions. This assess-infer process is, indeed, the essence of educational assessment.

Please look, then, at the three circles to the left in Figure 1.2 where you will see the most common decisions facing educators, that is, curricular, instructional, and evaluative decisions. A teacher's *curricular* decision might require an answer to a question such as, Which curricular aims should I pursue with this group of students? An *instructional* decision might hinge on answering the question, Do I need to make any adjustments in my ongoing instructional activities? And finally, an *evaluative* decision might call for an answer to a question such as, Should I change my just-completed instruction for my future students? Please note how you can see arrows from each circle indicating that, in order to arrive at a decision, the assess-infer process must be implemented. At the conclusion of this assess-infer process, the assessment-based inferences then contribute to the kinds of resultant decisions seen in the ovals at the right of Figure 1.2.

Clearly, there will be differences in the way educational assessment's two-step essence is applied in connection with the three kinds of decisions. For instance, when trying to choose curricular aims, the assessment to be employed would usually be regarded as a *preassessment* (or pretest) whose purpose is to get a fix on incoming students' entry knowledge, skills, or affective status. But when using the assessment

process in order to make en route instructional decisions, a variety of short-duration and informal assessments might supply the evidence from which the teacher could arrive at appropriate inferences about students' progress. And, finally, for evaluative decisions about whether to make changes in a just-completed set of instructional activities—ones that will be offered in the future to new batches of students—a comprehensive and formal assessment approach may be warranted featuring the use of more expansive posttests.

But, as Figure 1.2 portrays the situation, no matter what sort of educational decision is to be made, the use of educational assessment's essence—that is, its two-step assess-and-infer process—provides the evidence-based inferences that allow educators to arrive at the decisions most apt to benefit students.

Unwarranted Perceptions of Precision

Okay, now you've seen that educational assessment is essentially an inference-based tool to be used in the pursuit of improved educational decisions. This is a particularly profound truth—too often overlooked. Yet, once we have identified the proper use to which any tool should be put, this does not automatically indicate that the properly used tool is a terrific one. Even screwdrivers differ in their quality. The same is true of educational assessments.

If you'll agree that the underlying reason we assess students is to make better assessment-informed decisions about how to educate these students, there's a serious trap that, lurking out there, must be adroitly sidestepped. You need to dodge a perception that's held by far too many educators, namely, the fiercely flawed notion that educational tests are inordinately accurate. They aren't.

For the next few paragraphs, I'm going to be dealing with large-scale tests, such as the annual state-administered accountability assessments and the nationally standardized

achievement tests educators have used for almost a century. These large-scale tests are the sorts of measurement devices most people—educators and noneducators alike—think simply ooze accuracy. In contrast, when we think of the classroom tests constructed by teachers, most people recognize that such teacher-made tests can vary substantially in their quality—and that certain classroom tests are surely substandard. But the "big tests" are usually thought to be particularly precise measuring sticks.

It's easy to understand why many educators and most citizens believe large-scale educational assessments are remarkably precise. After all, these important tests are developed by measurement organizations that have been doing this kind of test-construction work for ages. Indeed, the first U.S. nationally standardized achievement tests were published as far back as the early 1920s. You'd think, after more than eight decades of creating such tests, by now the folks who make these assessments would assuredly have gotten it right.

Then, too, there is the ways in which the results of educational tests are usually reported, that is, in a numerical—sometimes remarkably opaque—manner that simply reeks of precision. Indeed, today's test results are sometimes reported as numerical scores *with decimals.* What right-thinking person would dare challenge a decimal-anointed test score? Test results from large-scale tests seem so, well, precise!

Finally, because it is rare for the accuracy of educational assessments to be publically challenged except, perhaps, when a company scoring large-scale tests makes a major mistake, almost all people—including most educators—ascribe more accuracy and meaningfulness to today's test scores than is actually warranted. Thus, if you voice a position that educational test results are capable of yielding some seriously misleading results, you're apt to be seen as out of step with most of the populace. Widely held beliefs can sometimes entice new converts, and a belief that large-scale tests are super precise is sometimes a consequence of the prevalent view that significant educational tests are, indeed, quite accurate.

Most school leaders have surely observed students responding to significant educational tests. Because, in most instances, these tests are administered, scored, and interpreted in a carefully structured *standardized* fashion, those school leaders might conclude that the test results emerging from this standardized system reek of precision. But what's not immediately apparent is that, because most significant educational tests are intended to tap students' mastery of large numbers of skills and substantial bodies of knowledge, the folks who create these tests are forced to rely on items that, at best, can only *sample* those skills and bodies of knowledge. As a consequence, depending on how the item-sampled content of a specific test happens to mesh with a given student's prior instruction and idiosyncratic background, in any given set of, say, 30 students, you can be certain that the results are apt to be equivocal for many students.

Wishing Won't Make It So

Assessment specialists sometimes get bad-mouthed for crimes they don't commit. To illustrate, one of the most impossible tasks we've given to those who must create educational accountability assessments is assessing students' mastery of too many curricular aims. Just as an airplane cannot fly simultaneously at several altitude levels—it's impossible—educational tests cannot properly measure students' mastery of too many curricular aims—it's equally impossible.

Yet when the curricular specialists of a state or province decide on the curricular aims to be promoted by teachers—hence the curricular aims to be assessed by accountability tests—these well-intentioned folks almost always identify too many curricular targets. In short, members of these high-level curriculum committees tend to choose the many things that they *wish* their students would be able to do. Such choices often lead to an extensive wish list of curricular aims—too many to be taught in any depth during the instructional time available and too many to be tested appropriately during the testing time available.

As a consequence, because test-constructors must rely on a *sampling* strategy whereby only some of the officially approved curricular aims will be measured on a given year's test, and not with enough items to yield a

(Continued)

(Continued)

meaningful fix on a student's mastery of any assessed curricular aim, test developers are berated by educators because of their unpredictable, sample-based accountability tests. But, in this instance, we should be faulting the curriculum blokes—a bunch of blokes who yearned for too much.

Then too, we must remember such oft-cited causes for potentially misleading test performances by students as day-of-test sickness, inadequate sleep, or family distractions. Most people realize that, on a given day, a student may perform much worse on a test than if the test had been taken on another day. And some kids, of course, are notoriously poor test-takers, especially when a test is thought by those students to be significant.

But these widely recognized reasons for less-than-perfect test taking reside in the students, not in the test. What is definitely not widely recognized is that the tests themselves yield imprecise scores. Later in the book, we'll consider what measurement specialists call the "standard error of measurement." It's a way of quantifying the amount of *imprecision* that's apt to be encountered in a particular test. In many tests, even really important tests, the standard error of measurement is quite substantial, thereby indicating there's apt to be a substantial amount of looseness in a student's score.

The implication of the imprecision of educational assessment instruments is that, although less than flawlessly accurate, those tests will usually yield evidence that, interpreted cautiously, *will certainly be far better than no evidence at all.* But school leaders should definitely not ascribe unwarranted accuracy to even the most carefully devised educational assessments.

Crucial Understandings

All right, we've come to the close of Chapter 1, and it's time to focus on the most significant understandings you should have snared as you read the chapter. As noted in the Preface, you

need to personally internalize these understandings sufficiently well so that, if necessary, you can explain them to others—particularly to your colleagues, to the parents of your students, to pertinent educational policymakers, and sometimes even to students themselves. So, after you've read a chapter's set of Crucial Understandings, please spend a few *extra* moments to consider how *you,* if you were required to do so, could get someone else to comprehend what's represented in each of those understandings. If you can successfully explain the nature of these understandings to others, you definitely will have understood what a school leader needs to understand about educational assessment.

A special collection of individuals to whom an astute school leader should give serious forethought are members of *the media.* Time and again, we see newspaper reporters who cover the local "education beat" truly mess up their stories about students' test performances. Distorted newspaper stories can, clearly, contaminate local policymakers' views about the effectiveness with which local educators are doing their jobs. So to minimize such distortions, school leaders need to remember one big-bopper lesson, namely, that *many media members who report on educational issues are astonishingly ignorant about educational assessment.*

Let me be candid. A newspaper's education beat is often given to journalists who are just getting under way with their careers, and we rarely see the reporters assigned to education stories choosing to remain for long in those posts. So, in general, we find a string of recent arrivals covering education events for local newspapers. And what these freshly minted education reporters usually know about educational testing will be what they recall from having taken tests while they were students themselves. So, unless a school leader gets remarkably lucky and finds a moxie, measurement-knowledgeable reporter covering local education stories, every school leader needs to do a proactive job in *educating* members of the media about any assessment-related issues germane to an upcoming story. Thus, as you find yourself trying to understand the notions treated in this book—and understand them well enough to explain those

ideas to others—remember that a key collection of those "others" are members of the media who will be reporting on local students' test performances. Just imagine that you are trying to explain each of the book's crucial understandings to a greenhorn reporter who barely knows how to spell *test.* If you can, think through what a reporter would most likely be interested in, and try to put together a reporter-friendly explanation of the assessment topic you're treating. Assessment-informed members of the media will, most of the time, do a far better job for school leaders than will media representatives who regard the results of educational tests as having been divinely derived.

For this chapter, there are two key understandings, and they're set forth below. Note the use of atypical type. If something is really *crucial,* don't you really think it definitely deserves a bit of atypical treatment?

Crucial Understandings

- Educators use assessment-elicited evidence about students' covert knowledge, skills, and affect to make inferences that can then contribute to more defensible educational decisions.
- Although it is widely believed that large-scale educational tests are remarkably accurate, they are much less precise than is thought.

Recommended Reading*

Dwyer, C. A. (Ed.). (2008). *The future of assessment: Shaping teaching and learning.* New York: Lawrence Erlbaum.

Reeves, D. (Ed.). (2007). *Ahead of the curve: The power of assessment to transform teaching and learning.* Bloomington, IN: Solution Tree.

Stobart, G. (2008). *Testing times: The uses and abuses of assessment.* New York: Routledge.

* Complete bibliographic information and brief annotations are supplied for the following recommendations in the Reading Recommendations Roundup (pp. 181–190).