

1

Why Fuss With Educational Testing?

This is a book about the basics of educational testing. But you might reasonably ask, why on earth would anyone actually wish to learn about the basics of educational testing? Answering this question is the focus of the book's first chapter. In this chapter I hope to convince you that many people, including *you*, really do need to understand the basics of educational testing.

But first, ever so briefly, let's look at what these "basics" of educational testing are. The title of this book, *The ABCs of Educational Testing*, was chosen because most people believe that the alphabet includes the fundamental building blocks of learning. Once children have mastered their ABCs, they're usually then capable of learning just about anything. What you'll learn in the ensuing pages is a handful of foundational concepts and procedures linked to the testing that routinely goes on in our schools.

Specialists in almost any field often assign distinctive labels to their field's concepts and procedures. Typically, those

labels are regarded as necessary because the concepts and procedures involved are being used by specialists in subtly *atypical* ways. Good intentions notwithstanding, the result of such distinctive labeling is that a specialization's content frequently ends up becoming essentially incomprehensible and even mysterious or off-putting to nonspecialists. The field of educational testing is no exception. I am not suggesting any willful camouflaging on the part of testing experts, only the understandable tendency of specialists to employ terminology that best reflects the nuances of their field.

Happily, the actual ABCs of educational testing, contrary to today's widely held view, are understandable *to anyone*. Moreover, there really aren't all that many educational-testing basics. Yes, the true ABCs of educational testing are relatively few in number and, even more importantly, can be readily comprehended by just about everyone. Accordingly, don't anticipate being baffled by the complexity of what you'll encounter in the following pages. The central concepts and procedures of educational testing are neither complicated nor excessively mathematical.

WHO IS THIS BOOK'S TARGET AUDIENCE?

By now you realize that a reader of this book is going to learn some basic stuff about educational testing, and learn it so well that—if asked—this reader can accurately explain this stuff to others. For whom, then, was this book actually written?

Five Target Audiences for a "Basics of Educational Testing" Book

- **Classroom Teachers.** Teachers need to understand how the quality of classroom tests can enhance—or can inhibit—their students' learning. Teachers also need to know how to discern whether the external tests they are required to use are in fact suitable for those tests' alleged functions.

- **Educational Administrators.** If educational administrators understand the contributions and limitations of educational tests, those administrators can better guide teachers regarding the appropriate uses not only of teacher-made tests but also about the appropriate roles for externally imposed exams.
- **Educational Policymakers.** If school board members and elected legislators understand whether particular tests yield evidence that's supportive of evaluative judgments about schooling, those policymakers can make more defensible decisions about the education provided in schools for which such policymakers are responsible.
- **Parents of School-Age Children.** Parents of children who are currently in school, or parents of younger children who will soon be in school, increasingly recognize that children's scores on educational tests can have a huge and sometimes lasting impact on decisions affecting children's in-school and beyond-school lives. Accordingly, parents may wish to know whether the educational tests used with their children are actually appropriate for making those decisions.
- **Everyday Citizens.** Because a society entrusts those who operate its schools with the responsibility to transmit and improve that society's culture, and because educational tests continue to play an important role in the way our schools operate, all citizens have both a right and a responsibility to see whether their tax-supported schools are performing successfully. To do so, citizens need to know if the evidence of school success based on students' test scores is accurate.

I tried to write *The ABCs of Educational Testing* for all five of these audiences. That's right, the book was written for classroom teachers, educational administrators, educational policymakers, parents of school-age children, and everyday citizens. How on earth did I acquire the audacity to tackle all five audiences simultaneously—in one single swipe? Let me tell you.

A few paragraphs ago, I argued that the most essential understandings regarding educational testing are relatively simple and, beyond that, are readily comprehensible to *anyone*. Well, I haven't altered that opinion already—after all, it was only a few paragraphs ago. I continue to believe that the ABCs of educational testing are understandable. What this belief translates into, therefore, is the conviction that if I can do a solid job in explaining what's meant by the key concepts and procedures of educational testing, those explanations can be grasped by members of any of the five potential target audiences identified above.

WHY THESE UNDERSTANDINGS?

If you look back for a moment at the five potential audiences for this book, odds are that you'll discover you fall in one or more of the five groups identified. You might be a parent of a fourth grader, a high-school English teacher, a member of a district school board, a city council member, or simply a citizen who cares about the well-being of our nation's schools. If you happen to fall outside those five groups, however, you might try pretending you're an educator or a parent as you complete the book. Extra motivation, even if mildly contrived, can often contribute to one's understanding.

A reasonable question you might ask yourself is, "So what?" Or, using other words, "Supposing that I do grasp the testing-related understandings about to be dished up in this book, what on earth would I ever be able to do with such understandings?" It is a reasonable question.

Well, the final chapter of this book will lay out a menu of potential actions that different readers of the book could take, and you'll definitely want to review those suggestions. However, even at this early point in the book, here are a few brief examples of possible action options available to members of each of the book's five target audiences.

Classroom teachers will be better able to generate their own tests to be used for particular purposes, such as improving ongoing instruction or evaluating the success of an entire semester's worth of instruction.

Educational administrators, such as school principals or assistant principals, can not only more skillfully support teachers' creation of effective classroom assessments, but school-site administrators can better discern which types of standardized tests are appropriate—or inappropriate—for evaluating a school's success.

Educational policymakers, for instance, an elected member of a suburban school board, after determining the degree to which students' performances on particular standardized achievement tests are truly indicative of how well the district's students are taught, can demand the use of tests that accurately evaluate instructional quality.

Parents of school-age children can first determine whether their children's teachers appear to use classroom tests in a way that's apt to enhance students' learning, and then—if necessary—support a teacher's greater use of classroom assessments designed to support instruction.

Everyday citizens can determine if media reports regarding students' performances on standardized tests are truly indicative of the degree to which those students were successfully taught and, if necessary, lobby for the use of more evaluatively accurate standardized tests.

In sum, the basic understandings promoted in this book are not merely “nice to know” for their own sake, but beyond that can serve as the springboards for readers to undertake a make-a-difference action. As indicated above, in Chapter 10 a batch of potential action options will be presented for the reader who—by that time—is most likely a *test-knowledgeable* reader.

CONVERSATIONS IN THE OFFING

It is often helpful to readers, particularly to readers of a possibly too technical book, if they know what's coming. Because the book deals with potentially complicated test-related topics, my treatment of those topics might easily become too technical for widespread understanding. Therefore, I have kept the book's language sublimely informal—much as if I were having a casual conversation with an educator in a school's Teachers' Lounge, or chatting with a passenger sitting next to me on an airplane. During such conversations, because I am supposed to know more about the book's content than most readers, I'll typically be the *explainer* while the reader is, I suppose, the *explained-to*.

To help you to get an early-on fix regarding your current capacity to understand many of the concepts and procedures treated in the following pages, you'll see on page 8 *A Confidence Inventory About Educational Assessment*. Please take a few moments to complete the inventory by anonymously registering the degree of confidence you *currently* possess if you were asked to undertake each of the ten activities described in this self-report inventory.

To avoid a potential terminology mix-up, however, before you tangle with the inventory, please recognize that in most of today's written materials about educational testing, the following labels are regarded as essentially interchangeable:

*Educational Testing = Educational
Assessment = Educational Measurement*

Almost everyone understands what is meant by “an educational test.” Yet, such tests (also called examinations or exams) are sometimes thought to consist only of the sorts of exams that most of today's adults experienced when they themselves were students in school—for example, tests incorporating multiple-choice, short answer, and essay items. To counteract such thinking, when referring to a “test,” many

writers prefer to employ the descriptive labels of “assessment” or “measurement,” which often bring to mind a wider variety of useful educational testing techniques, such as performance exams, oral quizzes, or collaborative problem-solving tasks. Although, at least for the present, the label educational “assessment” seems to be the most fashionable descriptor for educational testing, from this point on in the book, please regard the labels “testing,” “assessment,” and “measurement” as equivalent descriptors.

Now, with this terminology clarification in mind, please complete the self-report confidence inventory you’ll find on page 8. On page 9 following the inventory, you will find a brief scoring guide for the confidence inventory. After finishing the inventory, please give the scoring guide a brief look. When you have almost completed your reading of this book, you will again be provided with a copy of the confidence inventory so that you can once again complete the inventory to see if there have been any meaningful differences in your confidence level regarding certain aspects of educational assessment.

EDUCATIONAL ASSESSMENT**A Confidence Inventory About Educational Assessment**

VC = Very Confident	FC = Fairly Confident	LC = A Little Confident	NC = Not Confident at All
----------------------------	------------------------------	--------------------------------	----------------------------------

Directions: This inventory is intended to determine how confident you are with key educational assessment content. *Anonymously*, please indicate your level of confidence if *you* were asked to carry out each of the ten activities described in the inventory. Circle one of the following responses for each activity.

Suppose you were asked to	How confident would you be?			
1. describe to a family member what is meant by the label "student affect."	VC	FC	LC	NC
2. explain to a friend what the three chief purposes of educational testing are.	VC	FC	LC	NC
3. help settle an argument between two teachers by clarifying the difference between "validity" and "reliability."	VC	FC	LC	NC
4. make a brief oral presentation during a meeting of parents describing the key concepts of formative assessment.	VC	FC	LC	NC
5. describe to a parent how teachers should evaluate classroom tests intended to support teaching.	VC	FC	LC	NC
6. write a short note to a friend who is a state legislator in another state describing what is meant by "instructional sensitivity."	VC	FC	LC	NC
7. explain to a school's principal why there are different kinds of reliability evidence.	VC	FC	LC	NC
8. describe to a new acquaintance how today's concept of assessment validity differs from yesteryear's notion of assessment validity.	VC	FC	LC	NC
9. identify for a group of everyday citizens why it is the responsibility of those who design score reports to make them readily understandable.	VC	FC	LC	NC
10. explain to a newly elected member of a local school board how today's educational tests should incorporate assessment fairness.	VC	FC	LC	NC

A HALF-CENTURY MARCH TOWARD HIGH-STAKES TESTING

Educational testing has not always been such a big deal. Years ago, when I was a high-school teacher in Oregon, my students were obliged to complete a pair of nationally *standardized tests* every year. (Note: The glossary at the end of this book contains brief definitions of the book's terms identified in an *italicized and boldface* font.) Back then, however, annual standardized test-taking was mostly a ritual rather than an influencer of what went on in my school.

But this indifference to standardized testing evaporated in 1965 with the passage of the Elementary and Secondary Education Act (ESEA), a federal law of enormous significance. Whereas, prior to ESEA, almost all monies for public schools had come from state and local taxes, this groundbreaking 1965 statute supplied really serious federal dollars to state and local school systems. Although the vast majority of the funds needed to operate U.S. public schools were, despite the enactment of ESEA, still provided by state and local tax dollars, ESEA's new fiscal contributions to public schooling were greeted with elation by most state and local education officials.

CONFIDENCE INVENTORY INTERPRETATION GUIDE

This self-report confidence inventory is intended to help determine your perceived confidence in understanding a set of assessment-related concepts and procedures by securing your estimated confidence in explaining such content to others. To determine your total score on the inventory, simply assign the following per-item scores: VC = 3 points, FC = 2 points, SC = 1 point, and NC (or no response) = 0 points. Overall, then, the confidence you possess prior to reading this book can range from a high of 30 points to a low of zero. You may find it illuminating to recomplete the confidence inventory after reading the book. If you wish to retake the inventory, it is presented again on page 125.

Almost overnight, based exclusively on the evaluative requirements of 1965's ESEA, make-a-difference *educational evaluation* was born in the United States.

But there was a catch. Given the unprecedented nature of the federal funds provided via ESEA, many federal lawmakers were concerned that those funds might not be wisely spent by state-level and local education officials. Led by Robert F.

Kennedy, at that time the junior U.S. senator from New York, key requirements were inserted into the final version of the legislation so that state recipients of federal ESEA dollars were obliged to evaluate “this year’s” ESEA-supported educational programs in order to be eligible for “next year’s” federal largesse. Almost overnight, based exclusively on the evaluative requirements of 1965’s ESEA, make-a-difference *educational evaluation* was born in the United States.

Because it was generally conceded that the effectiveness of ESEA-supported educational programs should be evaluated chiefly according to how much students had learned, almost all early evaluations called for students’ scores on standardized tests to serve as the most important evaluative evidence. Because such nationally standardized tests as the *Iowa Tests of Basic Skills* or the *Stanford Achievement Tests* were usually sitting on the shelves in many district offices, and were widely thought to be credible measures of students’ learning, those tests soon became anointed as the chief tools for evaluating educational quality in our schools.

Relentlessly, an ESEA-fueled view of educational evaluation was accepted by more and more Americans—educators and noneducators alike—that students’ scores on standardized tests provided the best evidence regarding how well a group of students had been taught. Even today, it is a view widely held. As you will see in Chapter 4, however, it is a view that is often wrong—and harmfully wrong at that!

Because schools whose students performed below expectations on a state’s annual standardized test could receive serious federal sanctions, the function of such tests was to make educators responsible for the quality of their

instructional endeavors. As a result of this usage, these exams soon became known as *accountability tests*. For the most part, such annually administered accountability tests were either an off-the-shelf commercially produced standardized achievement test or, in some instances, a state-developed standardized test designed to measure students' mastery of a state's officially designated knowledge and skills. Students' performances on accountability tests, therefore, soon became the front-and-center freeway for evaluating the nation's schools.

In addition to standardized tests' evaluative function based on the performances of student *groups*, such tests also began being used to make important decisions about *individual students*. Given an increasing belief on the part of many U.S. citizens during the 1960s that some students—without being able to read, write, and compute—were still receiving high-school diplomas, widespread calls were heard for educators to measure students' minimum competencies. Indeed, students' passing a minimum-competency test soon became a high-school diploma hurdle in about half of our states. Students who failed to demonstrate, via their performances on such “minimum competency tests,” that they possessed at least rudimentary skills in reading, mathematics, and writing—after having been given several opportunities to retake a failed exam—were not granted a high-school diploma.

When we couple the negative consequences of a school's students' failing to score well on an annual ESEA-required accountability test with the diploma-denial function of high-school graduation tests in many of our states, it is understandable that these standardized tests were soon referred to as *high-stakes tests*. Unquestionably, they were.

But, laws can be revised, and during the more than fifty years that ESEA has exerted its powerful influence on American schooling, this federal law has been congressionally reauthorized a number of times. And, not surprisingly, in each of those revisions, changes—sometimes substantial—were

made. For instance, an influential reauthorization of ESEA in 2002 was the *No Child Left Behind Act* (NCLB), which carried with it a new series of daunting penalties for educators whose students failed to perform satisfactorily on annual state dispensed, but federally overseen, accountability tests. The most recent reauthorization of ESEA was a December 2015 bipartisan enactment of the *Every Student Succeeds Act* (ESSA). ESSA continued the NCLB requirement for annual testing of students at many grade levels, but assigned a substantial degree of oversight for implementing those assessment requirements to the states rather than to the federal government.

The *National Assessment of Educational Progress* (NAEP) is a congressionally authorized project of the U.S. Department of Education periodically assessing, on a sampling basis throughout the United States, students' performance in key subject areas. Because earlier administrations of NAEP revealed that different states had set markedly different levels of performance for students to be classified as satisfactory, in 2010 the federal government subsidized two state-level assessment consortia. Their mission was to create annually administered tests suitable for determining students' status with respect to the Common Core State Standards (CCSS), a set of curricular aims developed by two nongovernmental organizations intended for adoption by many, if not all, states.

Those assessment consortia, the Partnership for the Assessment of Readiness for College and Careers (PARCC) and the Smarter Balanced Assessment Consortia (SBAC), developed the necessary accountability tests suitable for measuring students' status with respect to the CCSS, but considerable backlash developed in various quarters of the nation against what some perceived to be a "federally imposed" national curriculum.

Because of political and parent opposition and other considerations, a number of states exited from their original membership in one of the two assessment consortia, preferring instead to develop—usually with external contractors' assistance—their own accountability tests. Membership in one of the consortia (PARCC) has dropped precipitously

low—including, at this writing, only a handful of states—while the other consortium (SBAC) has fifteen state members.

Given the uncertainty regarding how state education leaders will choose to implement ESSA in their own states, it is difficult to foretell how states' accountability tests will be established, and equally unclear about the assessment-related role to be played by the curricular aims embodied in the CCSS. Some states have simply adopted the CCSS as is, but used a different descriptor to label those goals. Other states have engaged in a substantial effort to identify suitable curricular targets and ways of assessing students' achievement of those targets.

THE FORMATIVE-ASSESSMENT PROCESS: INEXCUSABLY UNDERUSED

About twenty years ago a pair of British researchers, Paul Black and Dylan Wiliam, published a comprehensive review of roughly 250 sound research studies dealing with the *instructional* dividends of classroom assessment. Based on a meticulous review of studies involving teachers who employed classroom tests to improve students' learning, Black and Wiliam centered their attention on the use of classroom assessments to help teachers decide whether to make any adjustments in their ongoing instruction, or to help students decide whether they needed to make adjustments in the ways they were trying to learn things. This use of classroom assessment as an *instructional* illuminator is referred to as *formative assessment*. Black and Wiliam concluded "conclusively that formative assessment does improve learning" (Black & Wiliam, 1998, p. 61).

Students whose teachers employ the formative-assessment process almost always learn far more than do students in classes taught by teachers who fail to employ formative assessment.

The gains in student learning attributed to teachers' using the formative-assessment process are not only consistent, they are substantial. Students whose teachers employ the formative-assessment process almost always learn far more than do

students in classes taught by teachers who fail to employ formative assessment. As a consequence of the persistent and powerful payoffs of classroom formative assessment, it is apparent that the formative-assessment process should be more widely employed in our nation's classrooms. Yet, even though more pervasive use of formative assessment should clearly be encouraged, at the moment far too few of our nation's students are experiencing it. In Chapter 8 of this book, you'll see how formative assessment works.

TAKEAWAY TIME

Here's the most important understanding I hope you acquire as you read this chapter, complete with a handy label to help you remember it.

Twin Motivations for Assessment Knowledge: *Those who care about our schools should understand educational-assessment basics, not only because inappropriate tests often lead to mistaken high-stakes decisions but also because classroom formative assessment is being underused.*

Let's say that, while reading the chapter—often nodding emphatically in agreement as you did so—you internalized the above understanding. You can surely agree that there is nothing that's technically off-putting about this understanding. A straightforward translation might be that many folks need to understand educational testing because (1) we're frequently using the wrong tests to make key decisions and (2) we're not using formative assessment as much as we should. That two-part idea, or some paraphrased rendition of it, is what you should have gained from this opening chapter.