Online Supplement

For Group and Solo Readers of *The ABCs of Educational Testing: Demystifying the Tools That Shape Our Schools*

W. James Popham

CORWIN
A SAGE Publishing Company

Introduction

Not too many months ago, I sent off to Corwin, educational publisher nonpareil, the final touches of a book I had just written: *The ABCs of Educational Testing: Demystifying the Tools That Shape Our Schools.* The book is aimed at five target audiences, namely, teachers, school administrators, educational policymakers, parents of school-age children, and citizens in general. The purpose of my book is to help members of as many of those five audiences as possible become reasonably knowledgeable about a modest number of significant understandings regarding educational testing.

Chances are that you have already read all or part of what I'll hereafter refer to as *The ABCs.* However, if you haven't previously read all or some of that book, try to do so without delay. Reading *The ABCs* will position you to arrive at more defensible conclusions regarding how we currently use tests in our schools—and how we *should* use those tests. In *The ABCs,* I often explain (some might say *rant* about) the significance of educational testing's role in improving the way we educate our children. Indeed, as I frequently suggest in that book, when we consider the many alternative ways to enhance the quality of schooling, educational testing is, hands down, our most cost-effective way of doing so. That's right—improved decisions made on the basis of more appropriate uses of educational tests can emphatically boost the quality of instruction provided in our schools. And those improvements can be accomplished much more cost-effectively than other options.

But a nontrivial impediment to our getting the most mileage out of educational tests is that too few people really understand the fundamentals of what make such tests tick—or tock. Assessment-knowledgeable individuals—other factors being equal—will make better education-enhancing decisions than will individuals who are unfamiliar with the basics of educational assessment. Unfortunately, as matters stand, many of the individuals who now have a say in how our schools are run, including most educators, most policymakers, and most lawmakers, don't know enough about what today's education assessments can and can't do.

Accordingly, to encourage more folks to comprehend the assessment understandings treated in *The ABCs,* Corwin encouraged me to create this *Online Supplement* to that book. The *Online Supplement* was written in an unabashed effort to entice more people to read the original book and, thereafter, to better comprehend the assessment-related understandings presented in that book. The *Online Supplement,* although it can be used by an individual reader, should be of particular benefit to a group of readers such as a school's teachers, a school's parent organization members, or a district school board's members.

Organization of the *Online Supplement*

The ABCs contains ten chapters and, in the first nine of those chapters, presents nine significant assessment-related understandings. This *Online Supplement* is structured in a similar fashion. You'll encounter nine separate Chapter Extensions, each of which focuses on one of the understandings presented in the first nine chapters. A final Chapter Extension deals with the content presented in *The ABCs* book's last chapter.

Given that there are only nine understandings the organizational obstacles for group-study programs are not formidable. If, for instance, most of the language arts teachers in a large urban high school decide to study *The ABCs* during a school year, they might split up the nine understandings so

that, spaced over an academic year, 90-minute meetings every two or three months could focus on two or three chapters at a time. Ideally, of course, participants in such a group-study effort will have already read the group-assigned chapters before any book-focused sessions and will be in a position to deepen their understandings of a familiar topic during the interactive exchanges among a group's members.

Each of this *Online Supplement's* ten sections can be reached by clicking on the digital links provided herein.

Structure of the Chapter Extensions

The structure of each Chapter Extension will always be organized around the following four sections:

- **Better Understanding an Understanding.** After presenting a chapter's understanding, you will take a somewhat different look at the nature of this understanding than was given in *The ABCs.* Think of this initial section as an amplification or underscoring of a particular chapter's assessment-related understanding. Solo readers can enhance their grasp of a chapter's assessment-related understanding by mentally—or orally—trying to explain the meaning of a given chapter's understanding to another person—real or imaginary. In a group-based activity, such increased comprehension of a chapter's understanding typically takes place while comparing different people's interpretations of what a particular understanding actually means. The mission of this section of a Chapter Extensions is captured in its name, that is, to help readers in *understanding an understanding better.*
- **Collegial Conjecturing.** In this second section of each Chapter Extension, a fictitious colleague's e-mail regarding one or more features of the understanding is featured. In certain instances, the position taken by this e-mail writer will make slabs of sense. In other cases,

the fictitious writer will have authored a flock of foolishness. Your charge, either by yourself or in a group, will be to think through how you would respond to the make-believe e-mailer. If you carry out this "think-through" response in a group, of course, different members' responses can be shared with others. If you are conjuring up an e-mail response totally on your own, you can think through how you'd reply. Then, after a mental or oral presentation of your response (just to hear how it sounds), you may wish to applaud vigorously as a tribute to the wisdom and lucidity of your reply.

- **Thought-Provocation Queries.** More often than not, a thought-provoking question is preferable to a thought-suppressing question. And this is why many textbooks, particularly those written for reasonably mature students, often conclude their chapters with a collection of queries designated as "Discussion Questions" or some similar label. In the third section of each Chapter Extension, therefore, you'll find a small batch of queries intended to stir up, as Hercule Poirot would say, your brain's "little gray cells." In other words, to provoke cognitive consideration of the issues linked to that chapter's assessment-linked understanding. If you are using this *Online Supplement* by yourself, you could profitably try to frame, in your mind, how you might respond to each of the questions. If you are using the *Online Supplement* as part of a group, however, then different group members can be asked to respond aloud to the questions. Agreements or disagreements in the responses of group members can be discussed. The mission in such question-answering activities, of course, is to provoke more than a thimbleful of thought from you.
- **A Real-World Application.** Each Chapter Extension will be closed out by a description of an activity in which a group of participants will be asked to undertake an

endeavor as *if they had already comprehended the meaning of a particular chapter's assessment-related understanding*. The essence of this final group-structured exercise is to model the impact that such an understanding, if actually grasped and internalized, would have on one's real-world behavior. Although best undertaken when multiple participants are involved, a solo reader of *The ABCs* might profitably attempt to speculate about what differences, if any, would be seen between individuals *who had* or *who hadn't* grasped the full meaning of the particular assessment-related understanding being considered. Recognizing significant distinctions among people who have differing ideas about the implications of a given understanding can help people better comprehend an assessment understanding's essence as well as its likely impact.

To review, presented from here on in you will find in this *Online Supplement* a set of 10 Chapter Extensions, each of which contain the following sections: (1) *Better Understanding an Understanding,* (2) *Collegial Conjecturing,* (3) *Thought-Provocation Queries,* and (4) *A Real-World Application.* A reader who is employing the *Online Supplement* can choose to use none, some, or all of these four sections from any of the Chapter Extensions. Let's turn, then, to the first of our Chapter Extensions for, unsurprisingly, Chapter 1. Each Chapter's Extensions will begin with a boxed presentation of that chapter's assessment-related understanding.

1

Extensions

Why Fuss with Educational Testing?

Chapter 1's Assessment-Related Understanding

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

Better Understanding an Understanding

Although reasons for learning about educational measurement exist beyond the two motivations set forth in this chapter's understanding, those two reasons are particularly powerful in light of today's educational context. Indeed, either one of the two reasons—all by itself—should spur most people to learn more about educational testing.

Chapter 1's first reason for studying educational testing's basics—namely, to diminish the number of bone-headed decisions currently based on educational tests—is enormously important. In recent years, *literally* millions of children have been less well-educated because shoddy test-based policies have been installed and implemented. The most obvious instance of bone-headedness can be seen when the instructional quality of a school or a district is based chiefly on students' performances on "accountability" tests that have never been shown as suitable for such an evaluative mission. High scores on such tests are thought to be caused by successful instruction. Low scores indicate the opposite.

However, the tests were often developed exclusively to compare the performances of test-takers, and were never demonstrated *by any evidence whatsoever* to be capable of distinguishing between effectively taught and ineffectively taught students. Thus, such high-stakes accountability tests often supply a misleading estimate of a school's educational success—typically measuring what students bring to school in the way of differing affluence-determined experiences or inherited academic aptitudes. As a consequence, numerous errors are made in judging educational quality.

Many *effective* schools, because the wrong tests are being used, are evaluated negatively—so their staffs are urged to alter how students are being taught. On the other hand, many *ineffective* schools, because their students score well on the wrong kinds of accountability tests, are erroneously commended for their fine performance, hence urged to continue the dismal instructional practices they are currently employing. Who loses out in these far too frequent scenarios? Clearly, it is the kids.

Surely, many teachers and school administrators are being misjudged, and such misjudgments suck. Genuinely skilled teachers are not properly awarded, and truly inept teachers fail to receive the professional support they so desperately need. But when effective instructional procedures are jettisoned because of students' poor scores on the wrong tests, or

when shabby instruction is allowed to exist because of students' poor scores on the wrong tests, then it is all too apparent that the big losers are the kids themselves and the long-term loser is surely the society allowing such decisions to be made on the basis of students' scores on the wrong tests.

So, in looking at this initial chapter's assessment-related understanding, the first of two powerful reasons that you and others should learn about the fundamentals of educational testing is that misused tests, especially tests misused to make evaluative judgments about educational quality, definitely damage our children.

The second motivational reason embodied in Chapter 1's understanding is that we currently employ formative assessment in our schools far less frequently than we should be. Later in the book, Chapter 8 focuses exclusively on formative assessment. In that chapter, the nature and dividends of the formative-assessment process are addressed. Evidence, lots of it, that's supportive of formative assessment is trotted out for you. But for the time being, please believe that this potent *assessment-based* instructional procedure can provide huge educational benefits to students. Not to use it more frequently in our schools constitutes a whopping sin of omission.

So in a brief revisit to this chapter's assessment-related understanding, we see that we're (1) misusing certain educational tests and (2) underusing a potent assessment-rooted instructional strategy. Those two deficits, clearly, need to be corrected.

Collegial Conjecturing

Please review the contents of the fictitious e-mail below that was supposedly sent wafting through the ether to you by a colleague. If you wished to reply to the person who sent you the boxed e-mail, what would your electronic reply say? Remember, if you are carrying out this activity as part of a group, then different participants' responses can be

collaboratively considered by the group's members. If you are cruising through this *Online Supplement* solo, you will still find it useful to consider the *strengths* or *weaknesses* of how you might respond to your colleague.

TO: **A READER OF *THE ABCS FROM:***
A MAKE-BELIEVE COLLEAGUE (AND FAST FRIEND)
SUBJECT: **ENOUGH SCHOOL TESTING ALREADY!**

Hi:

I just had to write you today about several articles in our local newspapers—stories that I'm sure you'll have a reaction to. The articles all concern the excessive usage of educational tests in our schools. I know you've recently been thinking about the proper use of such tests, and I'm wondering whether you think what's being proposed by a nearby school district's board of education makes any sense.

Let me be more specific. Two weeks ago, our district's five-member school board *unanimously* transmitted a resolution to parents of the district's students asking those parents to have their children opt out of our state's annual standardized accountability tests. Those tests, in the resolution passed by the board, are seen as "too time-consuming, too costly, and of scant instructional value when educating children." The board's resolution even went on to say that if parents were inclined to do so, they could urge their children's teachers to dramatically reduce the amount of *classroom testing*—using teacher-made tests—because the total array of standardized tests so clearly gets in the way of students' learning.

So, what's your take on this board action? Does any of it make sense to you? Frankly, I find myself leaning toward the school board's stance because we often hear complaints from students about "too much testing." Please get back to me soon, because there's an open board meeting coming up on this resolution in two weeks. I plan to attend, and I don't want to look like a total ninny.

Thought-Provocation Queries

Please consider one or more of the following four queries related to this chapter's assessment-related understanding. Having done so, if the question(s) provoke even the tiniest bit of thought on your part, then consider how you might answer the question(s) you chose. Incidentally, I could have referred to these simply as "questions," but don't you agree that the use of "queries" adds a considerable touch of class to this activity?

Query 1. *The ABCs'* initial chapter tries to set forth a two-barreled rationale for someone to master a collection of understandings linked to educational testing. Do you think both barrels are really needed, or would *either* the misuse of high-stakes test results or the underutilization of the formative-assessment process be sufficiently motivating all by itself?

Query 2. Early on in *The ABCs,* it is stated that the book was written for the following target audiences: classroom teachers, educational administrators, education policymakers, parents of school-age children, and everyday citizens—in an undisguised attempt to help members of those five groups become more conversant with basic notions of educational testing. Do you believe it is possible to rank-order those five audiences according to which groups are more important to reach in this regard? If so, what would your rank-ordering look like (from the most important audience to the least important audience)?

Query 3. In Chapter 1's short history of U.S. educational testing, it was pointed out that so-called "accountability tests" became particularly influential after the enactment of federal legislation such as the Elementary and Secondary Education Act of 1965. Such statutes stipulated the annual use of standardized tests as one way of ascertaining how effectively the nation's children were being taught. Those federal requirements, because federal funds were being used to support specified educational activities, seemed quite reasonable to

most observers. However, what if the accountability tests being selected by state education officials were flat-out wrong? What if the chosen tests provided an inaccurate and misleading picture of educational success? How can well-intentioned educational laws, either federal or state statues, guard against inept implementation of those laws?

Query 4. Chapter 1 contends that substantial segments of our populace, particularly those segments of society most concerned with the quality of our schools, are remarkably ill-informed regarding educational assessment. Moreover, it is implied in Chapter 1 that many individuals *don't really know that they don't know* about educational testing. After all, most of today's adults completed a number of standardized or teacher-made tests during their own days as students, hence many adults consider themselves reasonably conversant with the ins and outs of educational testing. Do you think our nation's level of assessment literacy is really so low as seems to be implied in Chapter 1's assessment-related understanding? Why or why not?

A Real-World Application

This activity is intended for use in settings where a group of individuals is collaboratively digging deeper into the content of *The ABCs.* As indicated earlier in this *Online Supplement,* a reader who has no structured interactions with others readers of *The ABCs* might still benefit from mentally isolating the chief factors to be stressed during a collaborative implementation of the group-exercise *described below in italics.*

CONVERTING THE INCREDULOUS

(A GROUP EXERCISE)

The group's total members should be divided into two or more subgroups, each of whom is to take turns playing two roles. First, a

subgroup must plan and present a persuasive case intended to convince the undecided members of a particular school's Parent-Teacher Association (PTA) to initiate a serious, year-long self-study of educational testing. Second, a subgroup (perhaps along with other subgroups) must serve as an audience of teachers and parents who are being convinced to endorse such a study of educational testing.

As matters currently stand, many of this fictitious school's PTA members are uncertain about the merits of such a lengthy self-study commitment. Making the promotion of assessment literacy, that is, attempting to enhance teachers' and parents' knowledge about educational testing, will constitute a nontrivial commitment of both time and resources. Many teachers and parents, frankly, think that there are other more compelling topics for a year-long self-study effort of this sort. Each subgroup's task, therefore, is to devise an optimally convincing presentation to persuade other PTA members that a year-long focus on educational testing is what's most needed at this time.

Each subgroup should caucus for no more than 20 minutes to plan a panel-presentation of 5–10 minutes to others in the group. After the panel presentation is ready, then a subgroup actually presents its ideas to other participants in as compelling a fashion as possible. Thereupon, a 5–10 minute segment is allocated to questions or comments from the floor (that is, from members of the group who, at that moment, are playing the roles of other PTA members). Answers to these from-the-floor questions or responses to from-the-floor comments must be supplied by the panelists.

If time permits, it is usually helpful to spend a few "no-pretend" moments—having shed the activity-required roles of being panel presenters or members of a PTA audience—to constructively critique the quality of various subgroups' presentations. What is being fostered is an individual's grasp of the key requirements needed to persuade others to promote the assessment literacy. Although the fictional players in this exercise are parents and teachers, many of the arguments and persuasion ploys used in this setting will be applicable to other groups.

2

Extensions

Three Primary Purposes of Educational Testing

Chapter 2's Assessment-Related Understanding

> *Purposeful Educational Testing:* The construction and evaluation of educational tests should be profoundly influenced by one of the three primary purposes of such testing, namely, (1) comparisons among test-takers, (2) improvement of ongoing instruction and learning, or (3) evaluation of instruction.

Better Understanding an Understanding

The most important thing about this chapter's understanding is its effort to split educational testing's primary purposes into three mutually exclusive categories. If this division of educational assessment holds up, then all of us will have an easier time when thinking about the dominant measurement mission of a particular educational test. Please note that

two key assessment operations should be "profoundly influenced" by this three-way division among educational testing's primary purposes. Those two operations are the *construction* and *evaluation* of educational tests. That is, the way we build educational tests and—after they've been built—the way we measure how good they are. Given that the building and using of educational tests pretty much captures what's significant about educational testing, then this chapter's assessment-related understanding should be regarded as a significant one. A test's chief purpose, you see, should influence almost everything else in the educational assessment arena.

Chapter 2 discusses three chief uses of these tests, namely, (1) comparisons among test-takers, (2) improvement of ongoing instruction and learning, and (3) evaluation of instruction. There may be lesser purposes of educational assessments, such as satisfying some sort of governmental requirement to administer certain kinds of tests or, perhaps, satisfying the curiosity of a skeptical school-board member. And the virtues of these sorts of purposes should not be trivialized. They can sometimes be quite important. Yet, they are not a test's *primary* purpose. There is only one *primary* (most significant) purpose of an educational test.

Those who wish to become truly knowledgeable about educational testing must possess a crisp grasp of the tripartite distinction among purposes described in Chapter 2. One determination in the crispiness of one's grasp of these three primary purposes rests on an important distinction between two measurement notions, an *assessment purpose* and a *results application.*

Let's consider one way of distinguishing between assessment *purpose* and results *application.* An educational test whose primary purpose is to compare test-takers will, if the test is successful, provide scores that allow the performances of test-takers to be contrasted with one another. For instance, regardless of the kind of scoring system being used, one test-taker's score can be (1) the same as the scores of certain test-takers, (2) better than the scores of certain test-takers, or (3) worse than the scores of certain test-takers.

One of the most common ways of reporting student's performances on an educational test is to indicate how many score-points have been earned by a student as a "raw score" and then translate this raw score into a *percentile.* A percentile indicates what percentage of the test-takers were outscored by a particular student. For instance, a student whose raw score was equal to the 89th percentile, would have performed better than 89 percent of the other test-takers. If an educational test whose primary purpose is to make comparisons among test-takers produces a flock of different scores so such comparisons can be accurately made, then this test has fulfilled its primary purpose.

But how do we translate raw scores into *applications* leading to actual decisions? For example, we might rely on a set of percentiles to decide which students should be given awards for high-level mastery of whatever the test was measuring. Or, in contrast, we might use students' contrastable raw scores to identify a group of students who are then eligible for intensified remedial assistance because of their relatively low performances on a comparison-focused educational test.

Putting it differently, the test's *purpose* is to provide a set of contrastable performances by test-takers. This purpose is unarguably comparative in its intended use. However, the *applications* of the resultant comparative scores might vary substantially. All of those applications, however, hinge on this educational test's having accomplished its primary purpose, to provide performance-comparisons among test-takers.

Another primary purpose of educational testing is the improvement of ongoing instruction and learning. Let's see how a contrast between purpose and application plays out. If an educational test's primary purpose it to supply information that's usable in the improvement of instruction and learning, then such information must be provided in a consumer-usable fashion. Otherwise, if those who want to rely on a test's results can't actually employ those results, what good are such results in accomplishing a test's instructional-improvement purpose? For instance, let's say that a test has been constructed in such a way that the results it yields are at a "grain-size" that's far too broad to be of any practical

value—either to teachers who wish to improve their ongoing instruction or to students who wish to improve their ongoing learning. Because the test's primary purpose has not been accomplished, then its most important application, in this example the enhancement of instruction and learning, is destined to flop. Had the test's purpose been better satisfied—had more usable grain-size reporting of results been produced, then there would have been a more positive application result.

As a final illustration, let's consider the third primary purpose of educational testing, the evaluation of instruction. One application of tests with such a primary purpose is to identify truly terrific instructional programs so that those programs can be replicated elsewhere. Another obvious application of the results of an evaluatively focused test is to help root out ineffective instruction so it can be replaced with more effective instruction. However, a number of other applications of an educational test's results are surely possible. None of these applications of a test's results, however, are apt to be successful unless it has been demonstrated—with persuasive evidence—that the test itself is capable of distinguishing between well-taught and poorly-taught students. Putting it differently, any successful application of this third primary purpose of educational testing requires the provision of sufficient evidence that the test itself is *instructionally sensitive.*

Wrapping up this look at Chapter 2's assessment-related understanding from a somewhat different perspective, then, it becomes clear that those who would rely on the chapter's three-way division of educational testing's primary purposes need to keep distinct the overriding *purpose* of a test and the resultant *application* of the results yielded by the test.

Collegial Conjecturing

Please consider the contents of the brief, charmingly boxed e-mail presented below. It was supposedly sent to you by a close colleague. If you were in a mood to reply to this make-believe buddy, what would your own e-mail say?

Remember, if you are carrying out this activity as part of a group, then different participants' responses can be collaboratively examined by the group's members. If, however, you are sauntering through this *Online Supplement* solo, you can still find it useful to consider the *strengths* and *weaknesses* of your e-mail response to this colleague.

TO: **THE READER OF A BOOK ABOUT EDUCATIONAL TESTING** ***FROM:*** **A MAKE-BELIEVE CLOSE COLLEAGUE** ***SUBJECT:*** **WHEN IS TEST-USAGE NOT TEST-USAGE?**

Hi:

I know, from what you told me last weekend at Lester's party, you've been recently reading up a storm about educational testing. Why you would want to do such a thing escapes me. However, we should not let such silliness go to waste! Accordingly, I need your help.

A guy who works with me, Clyde, portrays himself as an expert—on just about everything. I think he is quite confused about one topic related to educational assessment, but I am not certain about how to handle him. Maybe you can draw on your recent reading to give me some ideas.

Clyde claims that, as he puts it, "A test is a test is a test." What he means is that if an educational test is *properly* constructed and *carefully* evaluated, then it can be effectively employed for a whole host of assessment purposes. For example, if a national standardized achievement test has been built by a major educational testing firm, odds are that this test can be properly used to support teachers' instruction, help school boards evaluate a school's instructional success, and compare the current achievement levels of students.

Clyde holds the view that well-made educational tests can successfully carry out a variety of missions. Do you think what he says makes sense? Is Clyde right or wrong? I'll be interacting with him later this week, and I'd really benefit from your thinking on this issue.

Thanks,
Lee

Thought-Provocation Queries

Look over one or more of the following three queries related to this chapter's assessment-related understanding. Having done so, if a query provokes even a smidgeon of thought on your part, then please consider how you might answer the query (or queries) you chose.

Query 1. Although we have experienced roughly a full century of large-scale, standardized testing in this nation, in recent decades the significance of the purposes to which large-scale educational tests have been used has become much more "societally significant." Which of the three primary purposes of educational testing treated in Chapter 2 do you regard as the test-usage most contributory to our current "high-stakes" testing? Why?

Query 2. If you were trying to explain to a friend why it is that the 2014 revision of the *Standards for Educational and Psychological Testing* (by AERA, APA, and NCME) have become so influential on the way today's educational tests are both built and appraised, what sort of explanation would you provide?

Query 3. It is argued in Chapter 2 that one of the most effective ways to isolate the primary purpose of an educational test is to identify the decision(s) riding on the test's results. Do you agree with this contention? Why or why not?

A Real-World Application

This particular Extension activity is less "real-world" than it is "pretend-world." However, as a group activity it can sharpen participants' conversance with a critical distinction between a test's *purpose* and the *application* of the test's results. As a group activity, it is intended to be used in settings where a collection of individuals is collaboratively probing the content of

The ABCs. As indicated earlier, a reader who has no structured interactions with others about the book can still benefit from mentally isolating the chief factors invoked during the sub-group exercise.

PURPOSE OR APPLICATION: WHICH ONE?

(A SUB-GROUP EXERCISE)

For this exercise, it is necessary to have at least two sub-groups, although additional sub-groups can be easily accommodated. The essence of the exercise is for members of each sub-group to caucus individually (and quietly), then work up descriptions of an educational assessment situation that clearly represents either a test's primary purpose or an application of a test's results. Ideally, each description will exemplify a purpose or an application, but will not be blindingly obvious to others. In other words, each sub-group might cause its colleagues to think harder about the content of this exercise by making their descriptions accurate but not embarrassingly blatant. Usually, 15–20 minutes is allocated for this activity.

To illustrate, suppose a group of a dozen parents is split into three sub-groups of four parents each. Sitting apart from one another, and keeping voices down, each sub-group then constructs three descriptions of either (1) a test's primary purpose or (2) an application of a test's results. The three descriptions can be all purpose-focused, all application-focused, or mixed. These descriptions should be accurate exemplars, therefore, of either test-purpose or results-application.

Each sub-group should generate its three descriptions so that they can subsequently be read aloud to the other sub-groups. Each description should be designated—in writing—by the originating sub-group as either a purpose or an application. (Think of this in-writing requirement as an obligatory answer key whose existence will tend to preclude belated answer key changes by those who contrived a description.)

Each sub-group then reads each of its descriptions, allowing other subgroups to indicate (as individuals or as a total

(Continued)

(Continued)

sub-group) whether they believe the description depicts a test's primary purpose or an application of a test's results. The originating sub-group then indicates which of the two options it was trying to illustrate. Disagreements can be briefly discussed.

This process is repeated until all sub-groups have had an opportunity to (1) present their descriptions, (2) have other sub-groups react to each description, and (3) discuss any disagreements. A general review of the examples, along with a reconsideration of the fundamental distinction between test-purpose and results-application can conclude this exercise.

3

Extensions

Behind Standardized Testing's Cloudy Curtain

Chapter 3's Assessment-Related Understanding

Standardized Test Development: Essentially identical to the procedures used when teacher-made classroom tests are built, the development of standardized assessments relies on particularly careful test-building and more complete—yet plain language explainable—documentation of purpose determination, content selection, and item construction/revision.

Better Understanding an Understanding

How to build and successfully use an interstellar rocket is a patently complicated enterprise. It's so complex, in fact, that very few people possess the smarts, training, and commitment necessary to pull off such an endeavor. Similarly, many people regard the standardized testing of our nation's students to

be an enterprise that's almost as off-putting as sending rockets into the dark beyond. As a consequence of the belief that standardized educational testing is somehow "beyond" them, too many individuals who should have a vital interest in the results of standardized tests simply avoid learning much about them, particularly how such tests came into existence. Most of today's citizens, who may have a rough ball-park notion of what standardized tests do, are frankly intimidated by the ways such tests were constructed and how they are then used.

Chapter 3 is written with an anti-intimidation mission clearly in mind. Today's standardized educational tests have become too important to leave them only to measurement specialists. Fortunately, as this chapter's assessment-related understanding asserts, the building of standardized educational tests is essentially no different from the way teachers crank out their own classroom tests every few days.

This chapter's understanding assumes that people who regard the creation of standardized educational tests as fundamentally unfathomable will usually be reluctant to look into the merits of such assessment instruments. On the other hand, those folks who recognize that the basics of standardized test-building are no different from the basics of test-building by classroom teachers will often be empowered to demand plain-language explanations of what went on when a specific standardized test was born. Such demands, though often warranted, are rarely registered.

Chapter 3's understanding indicates that three basic operations must always be undertaken when building any sort of educational test, whether the test is a nationally standardized achievement test or Miss Ballard's test of her students' mastery of punctuation. Those three test-development operations are (1) determination of the purpose for which a test is to be used, (2) selection of the knowledge or skills to be measured, and (3) generation of the test's items—and subsequent honing of those items if such honing is needed. This three-stage sequence will always be followed for any educational test worth its salt—or its pepper. It is always possible, of course, to

give short shrift to any of those three operations—but this treatment invariably reduces the quality of the resultant test. (Contrarily, there is no evidence that tall-shrift treatment of educational test appreciably improves them).

In the chapter's featured understanding, you'll find an important phrase that might easily be overlooked, namely, "plain-language explainable." It's a potent phrase because, once we realize that what's going on in the generation of a high-stakes educational test essentially travels the same three-stage trail that teachers follow as they gin up their own classroom tests, we can call for the creators of any significant standardized tests to spell out—*in language comprehensible to normal earthlings*—how each of the three fundamental steps of test-building actually took place.

Oh, sure, the explanations that are supplied of (1) purpose determination, (2) content selection, and (3) item construction/revision might to be dished up in language too technical for many. But an appropriate response from a listener at that point should be similar to a paraphrase of the following request: "Could you please explain that procedure again, but less technically?" If the person who's doing the explaining can't provide a clear and comprehensible, jargon-free explanation of how a test-development procedure took place, then the listener should accurately conclude that the explainer is simply not up to the job. *All* technical procedures associated with the building and sharpening of standardized educational tests can be explained in a way that regular individuals can intuitively comprehend. Moreover, once the nature of a test-development operation is understood, then the quality of that operation can be better judged.

To illustrate, in a July 24, 2016 opinion essay in the *New York Times,* Diane Ravitch, a prominent educational historian, renounced her previous endorsement of the Common Core State Standards, a set of curricular aims nurtured by the federal government. In her essay, Ravitch pointed out that the development of these Standards was funded "almost entirely" by the Bill and Melinda Gates Foundation. She also

characterized the development of them as "a rush job." Two sets of federally funded nationally standardized tests measuring students' mastery of the Common Core State Standards have since been developed and are now employed in various parts of the nation. Well, if you think back to the three basic test-building operations set forth in Chapter 2's test-related understanding, you will see that "content selection" definitely took place when the Common Core's curricular targets were identified. That particular phase of the Common Core's creation has never been adequately described to the public, but it certainly could be revealed if sufficient pressure to do so were present. The chapter's understanding indicates that all three stages of standardized test-building can be described in a transparent, comprehensible manner. All that's needed is for a sufficiently potent demand calling for such explanations to be present.

Building educational tests rests on common sense, whether such common sense is seen in the generation of teachers' classroom tests or in the construction of high-stakes standardized tests. Information regarding the quality with which the three chief test-building operations have been carried out can be, and should be, demanded by those who have a stake in the use of standardized educational tests.

Collegial Conjecturing

Now, please read the boxed, totally make-believe e-mail from an imaginary friend below. To add a spot of verisimilitude, we've called him Philip. Your task is to consider what your friend Philip is concerned about in his communication, then come up with a reasonable reply to him. Your response to Philip, if you are tackling these Chapter Extensions with others, can be compared to the replies of other individuals. If you are using this *Online Supplement* all by yourself, it can still help to think through what kind of reply you might whomp up for your fictitious friend.

TO:* A READER OF *THE ABCS FROM:
FERVENTLY FRIENDLY PHILIP *SUBJECT:*
OPACITY UNDER SCRUTINY

Howdy:

I am so pleased that you're back from your annual vacation. I hope you really enjoyed the islands. If you did not bring back gobs of photos from your snorkeling sessions, I'll never forgive you. However, I am writing for another reason than an appreciation of underwater photography.

You'll remember I told you that I've been taking an active part in the "Parents Together" group at the school where Fred and Fiona are now in the 3rd and 5th grade. Something came up during this month's meeting, only a week ago, and it has me really perplexed. I know you've been reading that book about educational testing that you told me about when we were together a few weeks ago. I can't recall the title, or even the author, but I think he was named Popoff—or something like that. Anyway, here's my dilemma.

A month ago, the governing Board of our 22-school district announced tentative plans to spend a huge amount of money on the purchase of what they call "standardized interim tests," a battery of three-tests-per-year for use in grades three through six. These tests are supposed to be administered several times during the school year, and the results are supposed to help teachers better mesh their instructional activities with the learning levels of their students. Because of the size of the district's investment (Read that as spending "Flo's and my tax dollars."), the Board is sending a representative from the company that wants to sell these tests to each of the district's schools. This person will make a visit and, if a school requests it, a second visit to each school. We had our first visit last month, and we've asked for a second visit from the same person in two weeks. Here's where you come in.

At the first of our two meetings, a Dr. Jill Havens of the testing company described the tests and how they were supposed to be used by our school's teachers. She did a solid job in laying out the potential uses of the tests, and most of our Parents Together

(Continued)

(Continued)

group understood what she was talking about. However, one of our members, Joe Simpson, asked Dr. Havens to explain how the tests were actually developed, that is, where they came from and how did they get to us? Joe wanted to know "how well the tests were developed." Dr. Havens replied that her company's standardized interim tests had been developed in accord with "best professional practices" and she was confident that they were, as she put it, "first-rate interim tests." At that point, however, I thought she became a bit condescending, and indicated that if we wanted more detailed information about the construction of the tests, she would need to return for a second meeting with us. She indicated, however, that she did not think most of us would understand the technically sophisticated procedures that had been employed to create these interim tests.

My question to you is the following: Can members of our parent group understand the test-development procedures employed with these standardized interim tests or, as Dr. Havens indicates, are those procedures too complicated for us? I really need your take on this issue. If she's correct, then we'll need to leave the "buy/don't buy" decision to others. But if she's wrong, then I am going to join Joe in demanding plain-talk descriptions of how these tests were built.

Thanks, Philip

Thought-Provocation Queries

Please take a gander at the following three queries, and select one or more to which you'd like to supply an answer. Then, if you're willing, churn up an answer to any of the questions that captured your fancy.

Query 1. In Chapter 3 of *The ABCs,* standardized tests are defined as follows: "A standardized test is a test that's administered, scored, and interpreted in a consistent, predetermined manner." As you can see, three required factors are identified,

the administration, scoring, and interpretation of a test's results. But what about the original construction of such a test? Do you think that different standardized tests can be built in meaningfully different ways? Why or why not?

Query 2. According to Chapter 3's assessment-related understanding, the description of a standardized educational test's development procedures must be "plain language explainable." However, procedures that are able to be explained in plain language need not always be explained in such a way. Or should they be? Do you think that plain-language descriptions of the procedures employed in the construction of a standardized educational test must, without exception, actually be provided? Or is the chapter's understanding satisfied when a plain-language explanation can be provided if necessary?

Query 3. Do you believe that the parallel drawn in Chapter 3 between the construction of teacher-made classroom tests and the construction of standardized tests—particularly the kinds of tests employed for high-stakes educational decisions—if accurate, really makes any difference? Or, instead, even if fashioned according to an identical test-development strategy, are the two kinds of educational tests so fundamentally different that such comparisons are of little utility?

A Real-World Application

The final Extension activity for this chapter revolves around the fact that standardized educational tests are merely more careful and better documented implementations of the same sort of test-development procedures that the world's teachers have been using since the middle ages and much earlier.

Please, in a group or by yourself, tackle the sub-group exercise described below in italics. You'll find that the need to come up with a rationale in support of Chapter 3's understanding will help you internalize it.

ENTICING LEGISLATIVE LEARNING

(A SUB-GROUP EXERCISE)

After dividing a larger group into subgroups of 5–7 members, imagine that each subgroup is charged with the task of testifying to the members of a state legislature's Select Committee appointed to study "The Role of High-Stakes Testing in Our State." More specifically, each subgroup is to describe why it is that the Select Committee's members should understand the basics of the way that high-stakes standardized tests are typically constructed and, after their development is finished, evaluated as to their quality. At least two of the legislators on the Select Committee have previously registered the opinion that the building and appraisal of educational tests—especially standardized tests being used for significant decisions—is a topic "best left to those measurement experts who really understand what's going on."

The task of each subgroup is to prepare an oral presentation of about five minutes' duration to be presented to the Select Committee. First discuss, then choose, what your sub-group regards as the most persuasive reasons for a state legislator to dip into a topic often regarded as too technical for laypeople—or even for educators.

After a preparation period of 10–15 minutes, have one or two members of your sub-group present to the total group your defense of the proposition embodied in this chapter's assessment-related understanding. Subgroups should take turns presenting their rationale in defense of the chapter's understanding. A final, full-group discussion regarding the strengths and weaknesses of different sub-groups' presentations should conclude this exercise.

4

Extensions

Validity: The Crux of the Caper

Chapter 4's Assessment-Related Understanding

> *Assessment Validation.* Assessment validation, the most significant process in all of educational testing, culminates in the creation of a validity argument based on evidence of both inference accuracy and the contribution of a test to the accomplishment of the purpose for which the test is used.

Better Understanding an Understanding

Though surely not as influential as The Ten Commandments—nor as old—the nine assessment-related understandings presented in *The ABCs* do, indeed, constitute an unarguably important collection of truths regarding educational testing. Thus, when Chapter 4's assessment-related understanding characterizes assessment validation as "the most significant process in all of educational testing," we should probably

pay attention to a chapter-understanding that regards itself as focusing on a "most significant process."

This is the first insight regarding the validity chapter's assessment-related understanding. It is such a truly important measurement understanding that a failure to understand it is almost certain to distort a reader's comprehension of *The ABC's'* remaining understandings. Yes, assessment validity is that important! Those who *snooze* in mastering this chapter's assessment-related understanding are definitely destined to *lose.*

The essence of measurement validity is succinctly captured in a single sentence drawn from the 2014 joint *Standards* of AERA, APA, and NCME. "Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (page 11). In that definitional statement, assessment validity reflects the extent to which evidence and theory are supportive not only of score-based *interpretations* (or, if you prefer, inferences) but also of the *proposed uses* (or, if you prefer, purposes) of tests. This, then, is the defining essence of educational measurement. We get students to display overt responses to certain sorts of assessment stimuli (such as a test's items) then arrive at hopefully accurate interpretations about those students' covert knowledge and/or skills. But the validation train doesn't stop there, it heads on to the issue of whether the test-based inferences we've drawn about students are truly supportive of whatever purpose the test is attempting to fulfill.

You will note that accuracy determinations of score-based inferences and a test's purpose-support are both *judgmentally* determined. There's not a clever, quantitatively determined number that allows us to definitively assign an A-plus or an F-minus to assessment validity when using an educational test. And this is why the establishment of assessment validity boils down to the use of a *validity argument* whose function is to help a test's users arrive at a defensible judgment about the degree to which a test yields accurate score-based interpretations *and* the extent to which those

interpretations are genuinely supportive of whatever a test's primary purpose is.

The centrality of purpose-support, particularly the presentation of evidence attesting to such support, reminds us of the need to have mastered Chapter 2's assessment-related understanding about the three primary purposes of educational tests. Clearly, whatever evidence is corralled for a test's contribution to decision-making must accurately match a test's dominant purpose. Inference-accuracy, though necessary, takes us only half-way home in assessment's validation sweepstakes. It is also necessary to show that accurate test-based inferences contribute meaningfully to the attainment of an educational test's intended use.

Looking back three paragraphs to the definition of assessment validity in the joint *Standards,* you can see that test-based inferences and purpose-support are confirmed by "evidence and theory." *Please* realize that, in this context, "theory" is simply a ritzy way of explaining how certain variables (as represented by evidence) are related to one another. Consequently, in establishing either the adequacy of a test's interpretive accuracy or its support for an intended use, the emphasis is almost invariably on the amount and quality of *evidence* bearing on both of those issues.

You may recall that in Chapter 4, an attempt was made to dissuade you from ever uttering the phrase "a valid test." This may seem to represent a trifling distinction to you, but in most instances it is far from trivial. If you, and your associates, can *routinely* regard assessment validity as a judgmentally determined attribute of a score-based inference's accuracy *and* the supportiveness of that inference to achieving a test's intended use—and *not* as an attribute of a test itself—then you've surely crossed the assessment-validity finish line with flying colors.

Because assessment validity represents the most significant process in all of educational assessment, this chapter's test-related understanding is, one would think, the most important of all of *The ABCs'* nine understandings. Such thinking is definitely on the mark.

Collegial Conjecturing

After reviewing the contents of the flagrantly phony e-mail below, please generate a response. Then, if you are tangling with *The ABCs* along with other folks, it will be useful to compare different people's responses to the e-mail's writer. As always, the thrust of this Collegial Conjecturing activity is to deepen your understanding of the chapter's assessment-related understanding.

TO: **A READER OF** ***THE ABCS FROM:***
LEE SMITHSON ***SUBJECT:*** **A TEST'S VALIDITY**

Hello:

You may recall meeting a few weeks ago at the Iversons' fund-raising party. As we were chatting, you indicated that you were currently reading "a fascinating and marvelously written" book about the ins and outs of educational testing. I am hoping that you've finished the book by now and can toss a few insights my way regarding educational testing because I am dealing with a remarkably opinionated friend whose opinions about educational testing are making me grind my molars!

What my friend, Tracy, believes—and she typically broadcasts her beliefs for all to hear—is that all high-quality educational tests are valid—regardless of the measurement missions they are attempting to fulfill. I think you said just the opposite.

Here's the event that precipitated my disagreement with her. This year the state has allowed school districts to select their own standardized tests rather than adopt the state-developed tests that annually measure students' mastery of state-approved curricular aims. Our district has chosen a national standardized test that was created as a college entrance exam for high school students—along with a number of lower-grade exams recently developed by the firm distributing the national test. Tracy claims—or should I say *proclaims*—that any test deemed good enough to be used as an admissions test by prestigious colleges is, thereby, good enough to be used for any of the purposes our district's

leaders wish. She says, and I am quoting her accurately, "A properly constructed education test will be valid for just about all assessment needs."

Is Tracy on solid or squishy ground here? If I recall our brief conversation at the Iversons' party, you had been reading that educational tests must be accompanied by evidence regarding their accuracy and their support of a test's intended use. I think you referred to this second factor as "fitness for purpose." What you told me would seem to contradict Tracy's strongly held view.

Could you please straighten me out on this issue? I'll be seeing Tracy again in a few days, and I would love to have some ammunition to contradict her. On the other hand, perhaps she is correct. Help!

Lee

Thought-Provocation Queries

Please look over the following collection of four queries deliberately fashioned with the hope that they would engender a dash of cerebral activity related to this chapter's test-related understanding or, more generally, regarding the chapter's focus on assessment validity. If you find one or more questions that catalyze your cerebration, try to come up with a defensible response to one or more of the queries you selected.

Query 1. Although in previous editions of the AERA-APA-NCME joint *Standards,* the importance of an educational test's purpose was identified, only in the most recent 2014 revision of the joint *Standards* do we see a strong commitment to the assembly of *evidence* supporting a test's contribution to a specific purpose. As a consequence, since the publication of those recent assessment "commandments," we now see markedly increased attention to whether actual evidence—not mere intent—confirms that a test is "fit for purpose." Why do you suppose this shift has occurred?

Query 2. Educational tests are sometimes intended to serve more than one purpose. To illustrate, certain state laws specifically indicate that annual state-sponsored examinations permit evaluations of schools' quality as well as make contributions to improved instruction. How should the *designers* of high-stakes, standardized educational tests deal with such a multi-purpose mandate? How should the *evaluators* of high-stakes, standardized educational tests deal with such a multi-purpose mandate?

Query 3. Why, in Chapter 4, do you think such a big deal was made of the perils flowing from reliance on "a valid test" conception of assessment validity? Is this merely a terminology affectation on the part of measurement devotees as they cavort in educational assessment's vineyards, or do you think it really makes a difference whether we say "a valid test" versus "validity based on evidence of inference-accuracy and purpose-support?"

Query 4. How significant is "most significant?" Should Chapter 4's assertion that assessment validation is "the most significant process in all of educational testing" be taken literally, or is this merely an attempt to have readers pay serious attention to this chapter's understanding? If you were attempting to use an educational test for one of the three major assessment purposes identified in Chapter 2, namely, to compare, to support instruction/learning, or to evaluate instruction, yet insufficient evidence was at hand to support a persuasive validation argument, how would you try to cope with such a situation?

A Real-World Application

As part of the following activity, you will be asked to imagine two sets of evidence for the two twin foci of assessment validation, namely, accuracy of score-based interpretation and

supportiveness of a test's intended use. Although set up as an activity to be carried out by pairs of participants, a solo reader of *The ABCs* can usually profit from undertaking what's asked of participant pairs in this exercise.

WHICH EVIDENCE WINS?

(A PARTICIPANT-PAIRS EXERCISE)

In this exercise, participants are to tackle the assigned activity in pairs. However, if three participants are involved, the exercise can still be carried out by a trio.

The focus of the exercise is the strength of evidence that can be employed to support either (1) the accuracy of a test's score-based inferences or (2) the supportiveness of a test's results in attaining the primary purpose for which a test is being used. Having split into pairs, each participant is allowed up to 10 minutes to arrive individually at two evidence descriptions that can be employed by potential users of the test to reach conclusions about assessment validity as it relates to the test. Although both of these descriptions (again, one emphasizing the accuracy of a score-based inference and one stressing the support supplied by the test for the test's chief purpose) contain positive evidence, one of the two sets of evidence is decisively stronger than the other.

After sufficient time has passed for individuals to isolate fictitious evidence of both kinds, one member of a pair describes (perhaps reads aloud from notes) the two sets of evidence—one dealing with inference accuracy and one dealing with purpose support. The other member of a pair indicates which of the two sets of described evidence is the more compelling—then learns whether this choice coincides with the originator's intent. At that point the roles are reversed, and once more the two sets of evidence are described and a decision is registered regarding the strength of the two sets of evidence.

If time permits, when all pairs (or almost all pairs) have finished their exchange of evidence sets, a general discussion regarding evidentiary strengths and weaknesses should follow.

5

Extensions

Reliability: Considering Consistency in Inconsistent Containers

Chapter 5's Assessment-Related Understanding

> *Assessment Reliability.* Assessment reliability, the consistency with which an educational test measures what it's measuring, is indicated by three conceptually different kinds of evidence and can be calculated either for test-taker groups or for individual test-takers.

Better Understanding an Understanding

Two significant yet distinctive notions are embodied in Chapter 5's assessment-related understanding. After acknowledging that "assessment reliability" equals a test's consistency,

it is then explained that a test's degree of consistency can be displayed in three fundamentally different ways. A follow-up notion is that an educational test's consistency can be represented either for *groups* of test-takers or for *individual* test-takers. We will soon take a closer look at the three ways of indicating a test's consistency—as well as the difference between a group versus an individual focus for determining such consistency.

But first, please note that reliability, unlike validity, is ascertained for an educational *test itself.* That is, we determine how consistently a test is measuring whatever it aims to measure. Whereas validity is not an attribute of the test itself, but describes the accuracy of a score-based interpretation and how well a test's results show attainment of an intended purpose. This reliability centers on the consistency with which the *test itself* measures.

However, as indicated in the chapter's assessment-related understanding, three fundamentally different conceptions of consistency have been traditionally employed by educational measurement specialists. It is important that you not only know how these three sorts of reliability evidence are determined but also are constantly on-guard whenever anyone trots out "reliability evidence." Do not assume that the sometimes cavalier designations of a test's reliability coincide with the particular kind of reliability in which you are most interested. More often than you'd suspect, such an assumption is unwarranted.

The three sorts of reliability-designation approaches are based on (1) test-retest evidence, (2) alternate-form evidence, and (3) internal consistency evidence. Because the first two of these require a pair of test-administrations, and the third (internal consistency) necessitates only a single test administration, it should come as no surprise that we bump into internal consistency estimates of reliability much more frequently than its two reliability siblings. Yet, because internal consistency reliability coefficients only indicate the degree to

which a test's items are functioning in a similar fashion, it should be evident that such coefficients tell us nothing about whether a test possesses test-retest reliability or alternate-form reliability.

Clearly, there is a profound practical difference between estimates of a test's reliability that centers on groups of students (such as all the fifth-graders in a school district) and a reliability estimate that focuses on a test's reliability for a particular fifth-grader—let's call her Tamara. When we lump together a large number of scores and consider their consistency as a coalesced group, the reliability of those scores will be decisively greater than what we calculate specifically for Tamara. Yet, teachers are often obliged to rely on a test's results to arrive at instructional decisions for Tamara, Tommy, and other *individual* students. Educational tests, when influencing decisions about particular students, usually measure with far less consistency than is widely thought. Those who use educational tests to supply results for individual students must become thoroughly conversant with what's embodied in a standard error of measurement (SEM). Interestingly, because the calculation of an SEM hinges on the type of reliability coefficient one incorporates into the SEM formula, the size of an SEM will be determined by which of the three types of reliability coefficients is chosen.

Collegial Conjecturing

Take a look at the e-mail below. It represents the sort of missive that a reader of *The ABCs* might receive from a close friend or even from a casual acquaintance. If you read what the hypothetical sender of this e-mail has to say, and still have an inclination to do so, please conjure up a response. The activity will help you more fully internalize the two chief elements of the chapter's assessment-related understanding.

TO: **THE READER OF A SUPPOSEDLY ENTHRALLING BOOK ABOUT EDUCATIONAL TESTING** ***FROM:*** **YOUR LONG-TIME PAL** ***SUBJECT:*** **IDEA VETTING**

Hello again:

I bet you didn't expect to hear from me again after last week's e-mail exchange, but something came up yesterday that you can make better sense of than I. I say so because of what you were telling me about the new book you were reading about educational testing. And, after all, what are friends for?

There's a guy I work with who is pursuing a master's degree at Tipton University, and he's been taking at least two courses regarding educational assessment. Yesterday, as we were having lunch in the departmental lounge, this guy started talking about the reliability of educational tests, especially for individual students. What he said really shook me up, and I wanted your take on his main point. Does it sound reasonable to you—based on what you've been reading?

His main point was that in many instances, if not most, we are *inappropriately* calculating the reliability of tests we employ to make instructional decisions about individual learners. Apparently, teachers (or sometimes counselors if a school has one) calculate what's called a "standard error of measurement" that's symbolized by the letters SEM. This SEM lets teachers know how consistent a student's score is apt to be on a given test. However, because there are actually *three* different ways of calculating reliability for a test, the wrong kind of reliability estimate is often used in an SEM's computation.

My friend indicates that, in most instances, an SEM is based on how similar a test's items are, but teachers are often more interested in how consistently a test measures students' ability over an extended period of time. As a result of often using the wrong input for an SEM's calculation, teachers are likely to make mistakes in their instructional decisions.

Does this make sense to you, or is the guy misunderstanding something? If you think he is on-target, I'd like to hear some more of his views. If you think he is off, I can find other lunchtime friends.

See you next month at Frank's, and thanks for taking a crack at this topic.

Jamal

Thought-Provocation Queries

Please review the following five questions linked—sometime loosely—to this chapter's assessment-related understanding about reliability, and mentally consider how you might answer such a question

Query 1. Test specialists often assert that, "Assessment validity cannot be present if an educational test is unreliable, yet having a reliable test does not guarantee assessment validity." In practical terms, what does this assertion mean? Do you agree with the statement? And, after agreeing or disagreeing, why?

Query 2. In Chapter 5 of *The ABCs,* a position is taken that busy classroom teachers need not calculate indices of assessment reliability for any but their very most important classroom assessments. Do you agree or disagree with that stance? And, having done so, as usual, what's the reasoning behind your own position on this issue?

Query 3. Busy teachers would often like to know how reliable the scores earned by their students are on significant standardized tests are, and even on significant classroom tests. Yet, in many instances the only sort of evidence about assessment reliability is reported for *groups* rather than for *individual* students. Are there any meaningfully helpful insights provided to teachers by group-focused reliability indices such as, for instance, a standardized exam's test-retest reliability coefficient? How would you defend your point of view regarding this issue?

Query 4. When some measurement-moxie educators chat about a test's standard error of measurement (SEM), they point out that, "All SEMs are not equal." Those educators realize that the size of an SEM is heavily influenced by the magnitude of whatever correlation coefficient has been chosen in the calculation of a specific SEM. What factors influence the choice of a correlation indicator in an SEM computation?

Query 5. Two types of reliability evidence are obvious contributors to score-based decisions about both individual students and groups of students. One of the three kinds of reliability evidence may not be. How might one or more relatively common educational decisions be influenced, if at all, by each of the three types of reliability evidence? Are there any kinds of reliability evidence that don't contribute to educational decisions?

A Real-World Application

Because Chapter 5's assessment-related understanding features three fundamentally distinct kinds of evidence regarding a test's consistency, you need to be thoroughly conversant with those three ways of portraying how consistently a test measures what it is supposed to measure. Distinguishing among these three evidence-types is the focus of the activity described below.

TANGLING WITH RELIABILITY'S BLESSED TRINITY

(A SUB-GROUP EXERCISE)

This exercise calls for the creation of sub-groups, perhaps 4–8 members in each. The overall structure of the exercise requires each sub-group to prepare descriptions of evidence chiefly indicative of only one of the three types of reliability, that is, test-retest evidence, alternate-forms evidence, or internal consistency evidence. This phase of the exercise usually take 10–15 minutes. These descriptions are then read aloud, one at a time, to the rest of the group (other than the sub-group that generated the description). After each description has been read aloud, members of the total group indicate—by raised hands—which of the types of reliability evidence was described. Members of a sub-group do not "vote" on their own reliability descriptions.

After each description has been presented and the larger group has had a chance to respond, the presenting sub-group indicates which type of reliability evidence has been described. Ideally the sub-groups' descriptions of reliability evidence should be demonstrably correct, but at the same time should not be so terribly obvious that little, if any, conversance with reliability evidence is really required.

Finally, each participant in the large group votes (by secret ballot, if preferred) to identify the sub-group whose descriptions were accurately labeled without being too blatantly obvious. In other words, a sub-group's efforts are regarded as excellent if the resultant descriptions of reliability evidence are definitively one of the three types of reliability evidence—without embodying an "in your face" obviousness. What's sought in this exercise is a display of subtle discrimination among the three types of reliability evidence one might encounter when using educational tests.

6

Extensions

Fairness in Testing: A Long Time Coming

Chapter 6's Assessment-Related Understanding

> *Fairness in Testing.* Fairness in educational testing, now seen to be as important as validity and reliability in the construction and evaluation of tests, must be carefully documented—employing both judgmental and empirical procedures—to maximally minimize assessment bias.

Better Understanding an Understanding

This chapter's assessment-related understanding contains a 16-word phrase, properly punctuated by commas at either of its ends. This 16-worder is much more significant than it seems at first glance. Here then, suitably shorn of its commas, is this potentially impactful phrase that focuses on fairness in testing: "now seen to be as important as validity and reliability in the

construction and evaluation of tests." And why, you might ask, is this collection of 16 words so very important? This would be a good ask.

To discern why fairness in testing has recently become so important, you'll really need to check on the two commodities that fairness in testing now matches in importance, namely, *validity* and *reliability.* Validity and reliability have been the hands-down heavy hitters in educational testing for eons, so if fairness in testing is currently occupying a comparably lofty position, then it is a level of significance not to be ignored.

The assignment of significance to the notion of fairness in testing was not made by a small gaggle of educators or a few measurement mavens. No, the significance of fairness in testing flows from the 2014 *Standards for Educational and Psychological Testing.* Moreover, we can safely predict that fairness in testing will become *even more important* as time goes by. Not only was fairness in testing given its own chapter alongside validity and reliability chapters in the 2014 revision of the joint *Standards,* but fairness in testing is now touted as equally important to those more traditional measurement constructs.

Note, too, in Chapter 6's assessment-related understanding that fairness in assessment is to be accorded great significance in both the *construction* of educational tests as well as in the *evaluation* of those tests. Actually, in the chapter itself we are urged to employ fairness "from the initial building of a test all the way through its evaluation, administration, scoring, and interpretation." We can reasonably foresee an ever-increasing—across the board—attention to fairness in educational testing.

Finally, the chapter's assessment-related understanding identifies the two procedures currently employed to enhance fairness in testing, that is, *judgmental* and *empirical* approaches. Clearly, a person who fully grasps the meaning of Chapter 6's assessment-related understanding must become reasonably knowledgeable regarding how these two strategies attempt to minimize assessment bias. Several of this Chapter's Extensions deal with those two distinctive but complementary techniques for squeezing unfairness out of educational testing.

Collegial Conjecturing

Please consider what is being said below in the imagined e-mail from Chris, a long-time friend of yours. It deals with an assessment issue that often pops up whenever the topic of fairness in educational testing is being considered. After reviewing what your pal has written to you, please decide whether you agree or disagree with the major point made in the e-mail. Then, having landed positively or negatively on what your friend has written, try to generate (mentally or in writing) a response to Chris's electronic communication.

TO: **A READER OF *THE ABCS***
FROM: **CHRIS *SUBJECT:* WHO GOOFED?**

Hi:

I've been remiss in not writing to you for the past few weeks, but I've been staggeringly busy at work, and haven't had any time to relax or to contact friends.

I'm writing to get your reaction to a point of view about students' performances on significant educational tests—such as the state-wide "accountability" tests taken a month ago throughout our state. I remember that you were reading a new book about educational testing, and your reading seems to bear directly on an incident that came up just yesterday at a dinner party for the residents of our subdivision. One of my neighbors, Floyd Jones, was complaining bitterly about the quality of tests being dispensed by our state's Department of Education. He indicated that those tests are "flagrantly" biased against certain sorts of youngsters, particularly students of color and students from lower socioeconomic backgrounds. Floyd is the parent of three African-American children, all of whom are enrolled in our local schools. Here's how he backed up his contention that our state's tests are, as he said, "bristling with assessment bias."

As you may know, our state administers standardized tests in grades 3–10 each spring in mathematics and reading. Students'

(Continued)

(Continued)

performances on those tests are released 10 weeks later for every district and report key demographic strata such as gender, race, and family income (as reflected by whether children receive any state or federal funding for at-school lunch or breakfast). The test results for the last school year were made public about a week ago. At every grade level where the tests were given, black and brown students performed significantly lower than their white classmates. To illustrate, with very few exceptions state-wide, the average percent correct on their grade-level's tests was almost 10 percentage points higher for white students than the percent correct earned by both Hispanic and African-American students. Those achievement gaps garnered the most headlines when our local media reported on the state-test results.

Because these particular performance disparities were seen in almost every corner of the state, Floyd is convinced that *the tests themselves* are biased against children of color. Precisely the same differences between white and minority students seen at the state level were also definitely present in our own school district.

Floyd was telling all those who would listen that the school district's parents need to band together and demand that the district leadership make available descriptions of the procedures taken during the state tests' development to reduce assessment bias. Thereafter, if appropriate, a group of district parents can send state officials—all the way up to the governor—a formal request to replace the state's biased standardized tests.

Based on your reading in that book about testing, what do you think of Floyd's recommendations?

I'll really appreciate any time you can give to this. Floyd is such a straight shooter that I'd like to come up with a sensible response to his concerns. Thanks.

Chris

Thought-Provocation Queries

By looking over the following four queries in this section of the Chapter 6 Extensions, you can determine whether any

of them are of sufficient interest to warrant a response—written or only mental—from you. Remember, if you are tackling these activities along with others, comparing your responses with the responses of others can be useful in isolating, then clarifying, the nuances associated with each query.

Query 1. If, for financial or practical reasons, you were obliged to put all of your fairness-promoting efforts behind *either* an empirical or a judgmental bias-detection strategy, which one of those two would you choose? And, having made your decision, please wrestle with the obligatory *why* that underlay your choice.

Query 2. One of the reasons educational measurement specialists attempt to minimize assessment bias is that, when present, assessment bias contributes to "construct-irrelevant variance." If you were attempting to explain the nature of "construct-irrelevant variance," what would your explanation be? That is, try to fashion an accurate and understandable description of "construct-irrelevant variance" for a *layperson*—not an assessment-knowledgeable person.

Query 3. If, as Chapter 6 in *The ABCs* asserts, attention to the reduction of assessment bias should be present during the beginnings of test development up to and including the actual administration of an educational test, how can such attention realistically be fostered? And, if fostered, how can such attention be effectively documented?

Query 4. Given the long-standing attention to assessment validation and test reliability, what are some potentially effective ways of engendering more serious attention to the elimination of assessment bias? How can this be done, not only among those who build large-scale educational tests, but also among regular classroom teachers and school administrators?

A Real-World Application

Please form small sub-groups, then spend about 15 minutes discussing the nuts and bolts, that is, the detailed particulars, of *one* of the four bias-reduction options presented below. What your sub-group's members should be trying to do is come up with a really sound, flaw-free procedure capable of withstanding the slings and arrows of skeptics. If you are tackling these Chapter Extensions on your own, you might still follow the italicized directions in the exercise, but do so mentally—and just for you.

GETTING SPECIFIC: ONE OF FOUR BIAS-BLASTING PROCEDURES

(A SUB-GROUP EXERCISE)

If you are working with a collection of others on this activity, then you should join together with 4–8 participants. What your sub-group is being asked to do is come up with a step-by-step description of one *of the following four procedures—each of which calls for the application of either a judgmental or an empirical approach to minimizing assessment bias:*

Employing judgmental bias-reduction tactics while constructing an important educational test

Using empirical bias-reduction techniques while constructing a high-stakes educational exam

Utilizing judgmental bias-reduction ploys while evaluating the quality of a significant educational assessment

Relying exclusively on judgmental bias-reduction procedures while evaluating the worthiness of an important educational examination

After choosing one *of the above operations, in 10–15 minutes or so devise what your group believes is an essentially unassailable plan to accomplish either test-development or test-evaluation, then present your plan to the remainder of the larger group. Those listening to this presentation should provide constructive feedback regarding the described procedures.*

If time permits, the entire group should engage in a discussion regarding the possibility of arriving at a defensible mix of judgmental and empirical procedures when building and appraising important educational tests.

7

Extensions

Reporting Results: Where Rubber and Road Meet

Chapter 7's Assessment-Related Understanding

Score-Reporting Rudiments. Because test-based inferences about student groups or individual students are typically formulated from score reports, users of such reports should demand that the scores being reported are both easily interpretable and have been elicited by a test whose purpose coincides with those inferences.

Better Understanding an Understanding

This chapter's assessment-related understanding addresses only two points. First, a plea is made for the *ease of interpretability* of score reports. Second, those who interpret such reports are urged to make sure that report-based inferences about test-takers reflect *consonance with a test's intended purpose.* Although readers will find a flock of other useful score-reporting

information in Chapter 7, the chapter's assessment-related understanding hinges dominantly on those two points.

Chapter 7 focuses on standardized educational tests rather than the sorts of test teachers churn out for their own classes. Interestingly, most teacher-made tests are readily interpreted because, after all, teachers need to make sense out of the tests they build themselves. In most instances, teachers crank out classroom tests that are altogether consonant with whatever purpose a teacher has in mind. Where we start getting into trouble with respect to this chapter's assessment-related understanding is with standardized tests, especially the sorts of tests whose use contributes to high-stakes decisions about the students taking those tests and/or the educators preparing those students for such testing.

The "trouble" that we start getting into with standardized tests stems directly from the two issues captured in the chapter's assessment-related understanding, namely, (1) the clarity with which a test's results are explained and (2) the degree to which a test's score-based interpretations are in accord with the purpose for which a test was built and evaluated. Putting it simply, a great many standardized educational tests supply score reports that are remarkably difficult to interpret. Similarly, a great many standardized educational tests report their results in a manner that fails to mesh properly with a test's primary purpose. Clearly it is difficult, and sometimes impossible, to make serious sense out of a standardized test's unclear results. Similarly, if those results are being put to a use inconsistent with the measurement mission for which a test was built, then such a test will contribute little to improving the quality of schooling. (And this, remember, is what most people regard as the underlying reason for educational testing's existence.)

With respect to the interpretability of a standardized test's results, this is an issue that has been addressed with vigour by members of the educational-assessment community for more than a quarter of a century. Prominent measurement specialists, such as Ronald Hambleton of the University of Massachusetts,

have led a concerted effort not only to report the results of standardized tests *accurately,* but also *understandably.*

Nevertheless, the difficulties of simultaneously attaining accuracy and understandability in reports of standardized-test results are formidable. As you read in Chapter 7, most standardized educational tests' results these days are reported in the form of "scale scores." For a number of statistical reasons, such scale-score reporting makes statistical sense because it can present equally difficult challenges to different test-takers. However, all by themselves, the scale scores in score reports from standardized tests often make no sense at all. Unless those who craft the score-reporting procedures from standardized tests get much more inventive than many of their score-reporting predecessors, what we end up with for standardized educational tests is a collection of remarkably uninterpretable results.

The second concern embodied in the chapter's assessment-related understanding is the need for persuasive evidence that reported standardized test results are based on a test that is demonstrably *fit for purpose.* Remember, if an educational test has been created, then evaluated, with an instructional-enhancement purpose clearly in mind, but that test's results are being dished up as though they were suitable for the evaluation of schools or teachers, then the test's results are blatantly out of whack with the test's purpose. Users of a standardized test's reported results must be constantly vigilant in making sure that test-results match test-purpose.

Collegial Conjecturing

If you're willing, please give a bit of scrutiny to the fictitious e-mail that you'll find below. This is a supposed e-mail from a buddy asking you to comment on a testing-relevant issue he's recently encountered. Because your friend knows you have been voraciously ingesting the measurement contents of a nifty new book about the fundamentals of educational

testing, and because your opinion is apparently valued, please try to respond to your friend either as a formally composed reply or, if you are in a make-believe mood, only mentally.

TO: **MY FRIEND AND RECENT DABBLER IN EDUCATIONAL TESTING** ***FROM:*** **YOUR AMIGO** ***SUBJECT:*** **TEST-REPORT INCOMPREHENSIBILITY**

Buenos Dias:

I need your smarts or, putting it more plainly, I need your smarts about how the results of educational tests ought to be reported. I remember your telling me, sometime when we were last together, that you've been reading some sort of "can't put it down" book about educational tests, and something has come up with the parents of our local school that—based on your "newly acquired insights"—you can probably clarify. It has to do with the way that students' test results are reported to parents.

First off, you'll remember that almost a quarter of the kids in Jose's elementary school have Hispanic backgrounds. In fact, some of the families arrived in the district less than two years ago. Fortunately, the staff of Jose's school has made a serious effort to engage these Latino parents in a wide range of activities. Because our family is Hispanic-American, Imelda and I have been asked to take part in some of these activities. It was during one of these meetings last week that the issue bubbled up about which I need your thinking.

In our school district, on three occasions during the school year, students in grades 3–5 are given commercially developed standardized tests that the district refers to as "interim tests." Parents receive reports of their children's performances on these tests within a week after the tests have been taken. The problem is that none of the parents—particularly Hispanic-American parents—have any idea what to do with these test reports. When several of the Latino parents asked during last week's meeting what these standardized tests were for, our school's principal replied that, "They are to be used for instructional improvement at school and at home." After the meeting, as we were walking to our cars,

at least three of the other parents came to me and asked in various ways, "How can we use the results of these tests if we can't make sense out of them?"

And this is where I hope you can help. Can the results of standardized tests be made understandable to parents—even parents whose first language is not English? Or is the reporting of standardized tests' results simply something that parents, especially Latino parents, aren't supposed to understand?

I'll appreciate any insights you can toss my way.

Javier

Thought-Provocation Queries

Please consider this set of three questions intended to evoke several dollops of thought from you. If you find yourself enticed by one or more of these queries, please fashion—mentally or in writing—a response. Sharing your responses with other individuals is not formally prohibited by the United Nations' Charter, hence might be a smart thing to do.

Query 1. Why do you suppose that the educational measurement community has had such a difficult time in coming up with score-reporting procedures that can simultaneously provide accurate representations of a test-taker's performance, yet be compatible with the primary purpose for which a test was created?

Query 2. One of the concepts advocated for in the development and use of educational tests is that they cleave to the idea of "universal design." This testing approach requires attention to be given to assessment needs of *all* test-takers from the moment an educational test is first contemplated—all the way until that test is administered and its results interpreted. How, though, should this quest for fairness in testing be portrayed, if at all, when test results are reported? For

instance, if meaningful—but appropriate—accommodations were made in the design and delivery of a test for certain sub-groups of students, should those students' scores be accompanied by an explanation describing the nature of such accommodations?

Query 3. Chapter 7's assessment-related understanding calls for users of a test's score-reports, particularly reports from standardized testing, to make certain that a test's results were elicited by a test whose purpose coincides with the kinds of score-based inferences being drawn. In practical terms, if you were a parent of a student taking a high-stakes standardized test, how might you determine the degree to which a test's score-based interpretations coincide with the primary purpose for which the test was built?

A Real-World Application

Although this chapter's assessment-related understanding targets only two important insights, Chapter 7 did deal with a number of other test-reporting procedures. These procedures contribute to the likelihood of satisfying the recommendations of the chapter's understanding. For the sub-group activity (4–8 members) presented below, you and your colleagues are asked to generate 3–5 *reporting-related* statements such as the following example: "A score-report based on *percentiles* indicates the proportion of a group of test-takers who have out-performed the student whose percentile is being reported." These should be declarative sentences that are definitively correct or incorrect. (The preceding example, two sentences ago, would be incorrect because it describes the *inverse* of what a percentile actually tells us.)

After your sub-group creates a handful of such reporting-related statements, some accurate and some inaccurate, you should read each statement aloud to the remainder of the group, and then ask the individuals in the larger group to indicate whether each statement is *Right* or *Wrong.* Supply the

correct answer for each statement and, if necessary, defend you group's choice about the rightness or wrongness of the statement. Other groups should do the same.

REPORTING RIDDLES: RIGHT OR WRONG?

(A SUB-GROUP EXERCISE)

After having assembled in sub-groups of 4–8 members, each sub-group should generate a set of three to five statements about reporting of test-takers' results based on the contents of Chapter 7. Some of the statements should be accurate; some should be inaccurate. That is, certain statements should "sound like" accurate chapter-based assertions regarding the reporting of standardized tests' results, yet contradict what's presented in Chapter 7.

After allowing, say, 10–15 minutes for this preparation of "ammunition" for the exercise, sub-groups then take turns presenting their statements, one at a time but read each statement aloud twice, then those in the larger group (other than the statements-originating sub-group) should indicate whether each statement is Right or Wrong. The sub-group presenting the statements then indicates whether the statement was Right or Wrong. Discussion can follow any substantial disagreement about the rightness or wrongness of a given statement.

This exercise is entitled Reporting Riddles: Right or Wrong? The exercise, of course, could have been described as a True-or-False endeavor. However, by so doing, the exercise's name would have abandoned its alliterative allure.

8

Extensions

Formative Assessment: Instructional Magic?

Chapter 8's Assessment-Related Understanding

Formative Assessment. The formative-assessment process, a robust, research-ratified use of classroom-assessment evidence permitting teachers to adjust their instruction and/or students to adjust their learning tactics—although remarkably effective—is seriously underutilized.

Better Understanding an Understanding

This chapter's assessment-related understanding deals with one big idea, namely, that a remarkably effective classroom instructional strategy is being woefully underused. What's most important for you to comprehend about this understanding, is the nature of formative assessment itself. If you don't grasp what formative assessment is—and isn't—then

it's almost certain that you'll not truly understand the thrust of this Chapter 8 understanding.

True, the chapter parades out a compelling collection of research findings to support formative assessment, chiefly in the form of a widely cited 1998 review of empirical investigations by two British researchers—findings subsequently confirmed by other investigators. But to get the most goodness out of the Chapter 8 understanding, it is imperative for a reader to really acquire a firm grasp of formative assessment.

For openers, it is crucial to recognize that formative assessment is a *process* in which assessments—typically classroom tests—are used in a particular way. These classroom tests, employed by teachers while instruction is taking place, are usually aimed at fairly long-term curricular goals, thus requiring an instructional sequence lasting for a month or longer. Because students will be completing classroom assessments during an instructional sequence consonant with formative-assessment process, it is typically efficient for formatively focused teachers to tackle fairly significant curricular targets when using formative assessment.

Another key notion of formative assessment is that the process can be used to enhance the instructional-adjustment decisions of teachers and/*or* the learning-tactic adjustment decisions of students themselves. Ideally, we would hope to see more classrooms in which the formative-assessment process is employed by *both teachers and students.*

The chapter's lament about under-utilization of a demonstrably effective instructional process, hinges on a thorough understanding of the nature of formative assessment itself. Accordingly, please try to internalize the following definition of formative assessment provided by a well-intentioned writer some years ago:

> Formative assessment is a planned process in which assessment-elicited evidence of students' status is used by teachers to adjust their ongoing instructional procedures or by students to adjust their current learning tactics. (Popham, 2008)

Don't forget the need for *planning*, and don't forget that the entire enterprise revolves around teachers' and/or students' making *adjustment decisions* about what they are up to. Most of all, please recognize that this potent assessment-rooted instructional process helps children get a better education. That's why we need to use it more widely.

Collegial Conjecturing

Presented below is a pretend e-mail to you from a pretend friend of yours. Your task is to look over the message sent by your friend, then construct a reply. You can agree or disagree with your friend's point of view, so simply make sure that your response, whatever its nature, is well-reasoned and persuasive.

TO: **THE PERSON WHO'S READING THAT BOOK ABOUT ASSESSMENT BASICS**
FROM: **YOUR LONG-AGO CLASSMATE**
SUBJECT: **WHY NOT MORE WIDESPREAD?**

Good afternoon:

Who would have guessed that, many years ago when we were both students in the same elementary school, I would one day be e-mailing you to get your opinion about the nation's inability to install a high pay-off instructional approach in more of our schools? Yet, this is precisely what I am doing at the moment.

I remember when you telephoned me a few weeks ago, that you have been digging into a new book about educational testing and, as you said then, you were often surprised by what you were reading. Let me lay out what's currently bothering me, and ask you to supply an insight or two regarding how to fix what I regard as an untenable situation.

I am referring specifically to the nation's under-use of the formative-assessment process. As I am sure you know, from your

(Continued)

(Continued)

reading of *That ABC's Book* (and I may have the book's name wrong), formative assessment calls for teachers or students to routinely collect classroom-assessment evidence and then, on the basis of such evidence, decide whether (for teachers) to adjust their instruction or (for students) to adjust how they are trying to learn. I believe I understand the essence of how this formative-assessment stuff works. What I *do not* understand, however, is why formative assessment is not utilized more in our schools.

The formative-assessment process, regardless of minor variations in how it is carried out, apparently pays off big time in improved student learning. Why on earth, then, aren't many more teachers using it? Everyone wants to be successful, and if employing the formative-assessment process will, as the research evidence seems to make clear, increase a teacher's instructional success, why is formative assessment not being used by every teacher in our nation?

Very appreciative will I be if you can get back to me on this! And, of course, I'll see you next month—even if you don't respond to this note!

Adrian

Thought-Provocation Queries

Please examine the four questions below regarding different aspects of the formative-assessment process. If you conclude that any of these queries are of interest, try to come up with a sensible response.

Query 1. By far, the bulk of supporting empirical evidence regarding the positive effects of formative assessment stems from the use of *classroom assessments* rather than from the use of *large-scale standardized assessments.* Why do you think this is so? Could it be differences between the test-construction procedures used with those two categories of assessments, or could it stem from the manner in which the two types of assessments are typically used?

Query 2. In Chapter 8 of *The ABCs,* the formative-assessment process is described as being organized according to a "learning progression." If you were describing a learning progression to someone who knew nothing about such things, how would you go about doing so? Beyond describing *what* a learning progression is, also indicate *why* learning progressions are often touted as important elements in the formative-assessment process.

Query 3. Please recall Chapter 2's distinction among three primary purposes of educational tests: (1) to provide comparisons of test-takers, (2) to improve ongoing instruction and learning, and (3) to evaluate instruction. Clearly, when classroom assessments are being used formatively, they are intended to help improve ongoing instruction and learning. However, some writers have argued that tests used as part of the formative-assessment process could also make an important contribution to the evaluation of schools or teachers. Do you agree or disagree with this position and, either way, why do you hold this view?

Query 4. At a school district, state, or national level (your choice), what practical strategies and tactics could be *successfully* employed in an effort to expand the usage of the formative-assessment process in classrooms?

A Real-World Application

As suggested earlier, to meaningfully comprehend this chapter's assessment-related understanding, it is necessary to fully grasp the nature of the formative-assessment process itself. In this exercise, subgroups are asked to devise realistic descriptions of the instructional use of educational tests. Some of those descriptions should be completely consonant with the definition of formative assessment that was provided earlier in this chapter's extensions, whereas some descriptions should not. The latter descriptions, although they may seem

similar to formative assessment, should actually not be. The exercise, described below in italics, is intended to help participants sharpen their abilities to distinguish between accurate and inaccurate descriptions of formative assessment.

DEFINITION DERBY

(A SUB-GROUP OR PAIRS EXERCISE)

This exercise can be carried out in small groups of 4–8 persons or in pairs. The essence of the activity is to have a sub-group or one member of a pair generate descriptions of the sorts of assessment-related instructional situations that might be seen in the real world of schooling. These descriptions should be in total accord with the 2008 definition of the formative-assessment process previously presented (and, thus, should be regarded as bona fide descriptions of formative assessment). Other descriptions should "look like" but not match the definition's key elements.

After allowing 15 minutes or so for a sub-group (or a pair-member) to develop a small collection of such descriptions, those involved should take turns in reading aloud their supposed examples of formative assessment in action. The other sub-groups (or pair-member) then decide whether each description should be classified as Real or Phony. Differences of opinion should be hammered out, if possible, in light of the definition governing this exercise.

9

Extensions

Students' Affect: Underappreciated and Under-Measured

Chapter 9's Assessment-Related Understanding

> *Affective Assessment.* Because students' in-school affective dispositions can have a significant impact on students' success, both during school and after its conclusion, student affect should be more frequently assessed via anonymously completed self-report inventories.

Better Understanding an Understanding

Much like the preceding chapter, this chapter's assessment-related understanding represents a position of advocacy. Whereas Chapter 8 urged the more widespread use of formative assessment, this chapter's assessment-related understanding

attempts to foster the greater use of instruments intended to measure students' affect. In order to thoroughly comprehend this chapter's affect-focused understanding, it's necessary for someone to become more intimately familiar with the nature of affective assessment.

Several significant distinctions can be drawn between educational assessments aimed at measuring a student's cognitive capabilities and measuring students' affect. To clarify, a student's cognitive capabilities include the student's knowledge and skills, whereas a student's affect encompasses the student's attitudes, interests, or values. We should review a few of those differences because they contribute to our awareness of what's required to support the chapter's assessment-related understanding.

One of the most influential of these distinctions is that cognitive measurements call for students to display their *optimal* level of performance, while affective measures call for students to provide *honest responses* to questions so that adults can arrive at accurate inferences about students' affect. This difference leads to a marked contrast between the kinds of inferences we can legitimately draw from cognitive assessments and those we can legitimately draw from affective assessments.

We can most efficiently make inferences about students' affect based on their responses to affect-related statements or questions contained in a self-report inventory of some kind. More exotic and more costly methods exist to discern students' affect—such as using rooms with one-way mirrors and employing accomplices to stage an "inference-contributing situation" related to students' affective dispositions. But, as a practical matter, self-report inventories are clearly preferable to elaborate measurement procedures such as those calling for costly biologically rooted tactics. Nonetheless, it would be foolhardy to believe that, when completing self-report inventories, students will always dish up the unvarnished truth when asked about their interests, attitudes, or values. For a variety of reasons, even when completing a self-report inventory without supplying their names, students sometimes distort their responses. As a consequence, when collecting

self-reported affective data, it is always necessary to be satisfied with *group-focused inferences.* These would include inferences about the affective status of the group of students who completed an affective self-report inventory.

For cognitively oriented tests, given that we are attempting to make inferences about an individual student's knowledge and skills, we make an inference that the test-taker's performance is indicative of an individual student's best ability-level. Such an inference, of course, will occasionally be mistaken. Sometimes on cognitive tests, students perform *below* their actual ability-level. Many factors might lead to instances of under-performance. On the other hand, we almost never see students performing well *above* their ability-levels on cognitive tests. Nonetheless, in most settings educators are reasonably comfortable in using a student's performance in response to a set of test-items in order to arrive at an inference (or, if you prefer, an interpretation) about an individual student's cognitive achievement level.

For affective assessments, on the other hand, inferences must relate to a larger collection of students, such as all students in Mr. Higgins fourth-grade class this year. After administering an affective self-report inventory to his 28 fourth-graders, Mr. Higgins could analyze students' responses in order to discern what levels of affect were present in his class. Four items in the self-report inventory might be centered on the routine ambiance of Mr. Higgins' class—especially on the levels of noise and student-movement during class sessions. Most students, responding with total anonymity, might indicate that they were "very often" distracted from learning by the tumult typically present in class. Indeed, almost half of Mr. Higgins' students might mark a response-choice on the self-report inventory indicating that they were "Often unable to concentrate."

Given such a result, Mr. Higgins will arrive at an inference that there's too much noise and movement going on in his fourth-grade classroom. A subsequent group-focused inference should lead Mr. Higgins to make adjustments in how he maintains order.

One final point needs to be made about the use of self-report affective inventories, namely, that the level of their accuracy is certain to be influenced by respondents who deliberately "inflate" or "deflate" responses regarding their true affective status. Ideally, the number of too-positive responses will be matched by the number of too-negative responses. However, that "perfect match" situation is unlikely. Although most self-report affective inventories urge respondents to complete such inventories anonymously and honestly, we must always recall that self-report inventories are, indeed, *self*-report inventories. Such instruments should be used more frequently by teachers, but teachers need to realize that the group-focused interpretations made from students' responses are apt to be rather indicative rather than completely definitive.

Collegial Conjecturing

Please read the imaginary e-mail below. You are to consider your friend's stance regarding an important issue associated with the measurement of students' affect—in this instance, the appropriateness of assessing certain kinds of values. After reviewing your friend's position on this issue, try to formulate how you would respond to this request for advice.

_TO:_ A TWO-TIME READER OF AN EDUCATIONAL TESTING BOOK _FROM:_ YOUR BUDDY, WHO CAN'T BELIEVE IT _SUBJECT:_ ACCEPTABLE AND UNACCEPTABLE AFFECTIVE EDUCATION

Howdy:

In last week's e-mail, you said you've decided to *re-read* that book about educational testing (I forget its title). Wow, you only finished reading it a couple of weeks ago, and here you are going for Round Two. What on earth is up with you? Did you not understand what you read the first time, or have your recollection skills disappeared?

Anyway, since you will apparently soon be a revered authority on educational testing, I want to run a position by you that I'm intending to push when my committee and I meet next week with our local district's school board. We've been scheduled to meet with them for nearly three months, and we'd like to talk about educational measurement—more specifically, the district's policy regarding the measurement of our students' attitudes, interests, and values.

Two years ago, the school board endorsed a policy of encouraging the district's teachers to more frequently employ self-report inventories in their classes—completed by students anonymously and focused on students' *affect.* I fear that some of the teachers have run amuck in their implementation of this policy. I need you to tell me whether my stance on affective assessment is defensible enough to present to the board.

I realize there is educational value in having teachers gauge the affective dispositions of their students in order to arrive at actionable interpretations of students' attitudes, interests, and values. In the two years that the district's teachers have been giving their students anonymously completed attitude inventories, we've clearly seen positive signs regarding several important attitudes, such as students' general interest in learning. Nonetheless, there are no guidelines from the district delineating what sorts of affect should be measured or promoted by teachers. And, candidly, I believe several district teachers have gone way overboard because a number of teachers are both *measuring* and then *promoting* students' acceptance of particular political values. A handful of teachers, in fact, have pooled their efforts and developed self-report inventories aimed directly at whether a student's political leanings were toward the right or toward the left. We can safely assume that, if those students' self-reported values were inconsistent with a teacher's political preferences, then some teachers might attempt—perhaps subtly—to alter such values.

My position on this affective assessment program is quite simple. I want the school board to stop allowing teachers to measure or promote *any values at all* or, possibly, *any values other than those approved by 100 percent of our nation's citizens.* If teachers

(Continued)

(Continued)

want to promote their students' honesty or patriotism, I'm fine with teachers' promoting students' acceptance of diverse points of view. However, the current policy of leaving affective assessment, by neglect, to "teacher's choice" is, in my mind, quite dangerous.

Well, that's my position. Do think it has sufficient merit for me to voice it before the board? I know you've been thinking about this sort of measurement issue, and I really do need your advice! Let me know whether you believe that only values endorsed by *practically everyone* should be assessed and promoted in our public schools.

Les

Thought-Provocation Queries

Please consider the three questions below, each of which is related to affective assessment. Then, if one or more of the questions strike your fancy, think through how you might go about churning out a consummately compelling response to the question(s).

Query 1. Because most students have not experienced a great deal of affective assessment, particularly the completion of self-report assessment inventories taken with total anonymity, some students will need to be given an explanation/orientation regarding why this sort of assessment is taking place. If you were being asked to explain to a group of students the rationale for the assessment of their affective dispositions, what would be the chief elements of the explanation you might compose?

Query 2. A "cognitive laboratory" is an activity during which a small group of test-takers take part in an item-by-item analysis of an under-construction assessment device. Often used for cognitive assessments focused on measuring students'

knowledge and skills, cognitive labs can also be quite helpful in constructing self-report inventories to assess affective dispositions. If you were in charge of a cognitive lab to help improve the qualities of a new self-report inventory intended to measure students' interest in non-required reading, what would you focus on when interacting with a group of 6–10 seventh-grade students?

Query 3. If you were a fourth-grade teacher committed to measuring a small number of your students' affective dispositions over the course of a school year—once every few months—which *three* affective dispositions would you choose to measure via self-report inventories and address instructionally?

A Real-World Application

In the real-world measurement of students' affective dispositions, the most common assessment procedures used are self-report inventories anonymously completed by students. The phrasing of the items in such inventories is particularly important. Typically, a student is presented with a series of statements such as, "When I get home after a day in school, the very last thing I want to do is study." Students then choose a response to each item from options often ranging from Strongly Agree to Strongly Disagree. To make such inventories sufficiently efficient, the statements must elicit differing degrees of agreement or disagreement. For instance, almost all students would register disagreement with a blatantly positive statement such as, "I enjoy school so much that I desperately wish they held it seven days a week—with no weekends at all!"

Your task in this chapter's final Extension is to generate a set of three-to-five statements for an affective inventory intended to measure seventh-grade students' attitudes toward the subject of mathematics. Try to make the statements definitely positive or negative toward math, but not so flagrantly positive or flagrantly negative that there would be little variation in students' responses to an item. Details of the exercise are supplied below in italics.

CONSTRUCTING ITEMS FOR A SELF-REPORT INVENTORY

(A SUB-GROUP EXERCISE)

Because the mission of this exercise is to increase participants' familiarity, hence comfort, in creating the items needed when developing an affective self-report inventory, it will come as no surprise that the exercise calls for the generation of such items. First, please divide your larger group into sub-groups of 4–8 members, and then spend about 20 minutes writing items for an inventory intended to measure seventh-grade students' attitudes toward mathematics. Try to do a good job in generating the statements that could, if regarded as suitable, help constitute an early version of a self-report inventory.

The items in this inventory are usually statements relevant to the affective variable of interest—in this instance students' attitude toward mathematics. Try to create relatively short statements, using age-appropriate vocabulary, that represent either a person's positive or negative sentiment towards mathematics. An example of a positive statement might be: "I really have fun trying to solve math word-problems." An example of a negative statement might be: "When compared to my other subjects, I like math the least." In the inventory itself, students will be selecting one of the following responses for each item: Strongly Agree, Agree, Not Sure, Disagree, or Strongly Disagree.

The statements you generate should (1) help contribute to someone's drawing a group-focused inference regarding students' attitudes toward mathematics and (2) although definitely positive or negative, dare not be so positive or so negative that there is insufficient variation in students' responses to an item (that is, not enough variation to permit determination of a student group's affect). At the close of a sub-group's item-writing time, each sub-group's draft statements should be read aloud so that other participants can critique them

As usual, if someone is reading The ABCs solo, and thus has no opportunity to carry out this exercise as a member of a sub-group, the task of generating items for an inventory—and subsequently reviewing them—can be done individually.

10

Extensions

What's Next?

Introduction to a Conclusion

In the classic television drama, *The West Wing,* President Jed Bartlett often prods his advisors to move forward on new governmental initiatives by asking them, "What's next?" In a similar vein, *The ABCs* was written with the hope that at least some of those who read it would employ whatever they were learning to improve education. Accordingly, the final chapter of the book poses an overriding "What's next?" query to readers. That is, Chapter 10 entices readers to *do something* with their newly acquired insights regarding educational measurement.

Unlike the book's first nine chapters, each of which focuses on promoting the reader's mastery of a specific assessment-related understanding, Chapter 10 is organized into more of a "wrap-up" structure containing four separate "Chapter Chunks." These Chapter Chunks deal with (1) the complete set of nine per-chapter understandings, (2) a reprise of the self-administered confidence inventory first presented in the book's initial chapter, (3) potential understanding-triggered

actions on the part of readers, and (4) a small set of assessment-related issues that your affable author believes warrant special attention from readers.

Rather than trying to squeeze Chapter 10's decisively different content into the pattern used for the previous nine chapters where, for instance, there was always a section dealing with "Better Understanding an Understanding," the final Extensions for the book are organized quite differently. In the following pages you will find commentaries (all mine) on each of Chapter 10's chief segments. These comments, sometimes quite brief, are intended to help you reach a more defensible stance regarding what, if anything, you intend to do with the stuff you've been learning from *The ABCs.* The structure of this chapter's Extensions, therefore, will coincide completely with Chapter 10's four Chapter Chunks. Let's get into those four chunks.

Chapter Chunk 1: The Per-Chapter Understandings

What's first offered up in this chapter's Extensions is a complete set of the nine chapter understandings around which *The ABCs* is organized. Presented in the same order as they appeared in the book's nine chapters, and accompanied by a brief descriptive label for each, these nine assessment-related understandings clearly constitute the book's most important content. The nine understandings listed in this initial Chapter Chunk should, therefore, represent a reader's most important overall takeaway from *The ABCs.*

Let's be candid. As the writer of *The ABCs,* I decided to exercise an author's prerogative by laying out what I regard as the nine most important things about educational assessment that should be comprehended—and, hopefully, internalized—by members of the five target audiences for whom *The ABC's* was written. I believe that if a person actually comprehends these nine assessment-related understandings, then

such an individual would be—in my judgment—truly *assessment literate.*

If another member of the educational measurement community had set out to write a similar book about educational testing, would that author's list of assessment-related understanding coincide perfectly with my "gnarly nine" understandings? I doubt it. Nonetheless, I think there would be substantial overlap among most educational assessment specialists regarding what's truly important about educational testing for teachers, administrators, policymakers, parents of in-school kids, and regular citizens. So, although you should not regard the nine assessment-related understandings presented in Chapter Chunk 1 as representing holy writ, you can reasonably conclude that they deal with most of the important aspects associated with educational assessment.

Do you need to memorize the nine understandings in order to have truly comprehended them? I don't think so, there are better things to memorize. Instead of committing them to memory, grasp the essence of each assessment-related understanding, then cognitively internalize that understanding—*in your own words.* You really don't have to mimic the precise verbiage of the nine understandings as presented in *The ABCs.* Instead, you need to comprehend what's present in all nine understandings so that, if a meaningful educational issue meanders your way, you will be in a position to provide a dash of illumination to the resolution of that issue.

Chapter Chunk 2: A Reprise of the Confidence Inventory

Because I had promised in Chapter 1 of *The ABCs* that the self-report confidence inventory presented in that first chapter would be made available in the book's final chapter, and because publishers become visibly annoyed when authors brazenly double-cross their readers, that same confidence inventory is presented in Chapter 10. (I am not implying that

publishers countenance non-brazen, sufficiently sly double-crossing of readers.)

If you completed and scored the inventory when reading Chapter 1, I think you'll find it interesting to re-take the confidence inventory again to see if there's been any serious shift in your score. I am hoping, as you might suspect, that you will see a major increase in your assessment-related confidence. Remember, the items in the inventory typically call for you to register your confidence in being able to "explain" or "describe" (or some similar expository action) an educational testing concept or procedure. Consequently, it should follow that people who have comfortably grasped the meaning of *The ABCs'* nine assessment-related understandings would display greater confidence than would people who haven't.

Some readers, even those who have truly mastered the book's nine understandings, may not display an increased level of confidence between their completions of the inventory. Certain people are inherently not confident sorts of blokes. And this is why, as explained in Chapter 9's treatment of affective assessment, it was regarded as unwise to arrive at affective-focused inferences about *individuals* based on self-report inventories. However, if you happen to be dealing with *The ABCs* in a group, for instance, as a member of a school PTA association studying the book, administration of the confidence inventory on a pre- and post-basis ought to result in some serious increase in confidence scores. The resultant inferences about the *group's* affect would be far more valid than inferences based on the responses of individuals.

Chapter Chunk 3: Understanding-Triggered Action?

As confessed on several occasions in *The ABCs,* I wrote the book because I am convinced that appropriately created and properly employed educational tests represent the most cost-effective way of improving the education we provide to our

children. Yet, even though the book does a solid job of informing readers about a collection of assessment truths, what if *none* of those readers ever took any action regarding what they had read? Such a situation would be analogous to the oft-cited forest puzzle with its fabled falling tree that, if unheard, does or doesn't make any noise. Clearly, what I hope you'll do is take action, a little or a lot, to stimulate the better use of educational assessment to benefit more students.

I realize that it may seem laughable to expect many readers of *The ABCs* to wrap up their reading of the book, then sit down instantly to consider what steps they should undertake in an effort to better educate children. What seems more likely is for someone who has internalized *The ABCs'* nine understanding to be routinely "on the lookout" for assessment-related shortcomings that can be remedied. Then, when there's a tangible situation in which a better understanding of educational assessment's basics might improve what's going on—a reader is positioned to *get into it with gusto!* In other words, you can scan the scene to see if you believe what's going on should be applauded, changed, or stopped completely. The right question, voiced at the right time, can force participants to think more carefully about their current uses of educational tests.

In most situations, because they are now more knowledgeable about the contributions and limitations of educational assessments, many readers will prefer to be "at the ready" rather than "aggressively active." Clearly, the choice is yours.

What's particularly important in this context is for you *not* to assume that assessment-related flaws will be fixed by others—others who possess more assessment sophistication than you. Sadly, assessment literacy is not widespread in our nation. Indeed, few of today's teachers and administrators are familiar with the kinds of issues associated with *The ABCs'* nine assessment-related understandings. It's currently unlikely that many assessment-knowledgeable persons will be available to "fight the good assessment fight."

The collection of "potential action options" presented in Chapter 10 of *The ABCs*—options that might be taken by a person who wants to use assessment for improved education are, hopefully, illustrative. They most assuredly are not exhaustive. Thus, think hard and inventively about how you could employ your newly acquired insights regarding educational testing as a springboard for stimulating constructive change. You might embark on a change-promotion with abundant energy, or you might simply raise one penetrating question at an appropriate moment. Even a gadfly can sometimes cause serious thinking by others.

Chapter Chunk 4: Quick Takes on Educational Testing

Other than the Glossary and Index, *The ABCs* wraps up its story in Chapter 10, and thus presents the final opportunity for the book's author to blabber about important aspects of educational testing. I chose to address three assessment-related issues and, while getting ready to write this particular Chapter Chunk, I went back and re-read what I had previously written about those three issues in Chapter 10.

For this final Extension, I have no intention of re-stating what I already said about my three chosen issues. Accordingly, after I completed my re-reading of the fourth and final Chapter Chunk in Chapter 10, I decided to crank out one paragraph only as a re-think regarding the issue being treated. My hope is that you'll look over those three paragraphs, then decide whether you regard my three solo paragraphs as sensible.

Assessment literacy. I've devoted considerable thought to the ways we currently use and misuse educational tests. I could have said "I'd spent many sleepless nights worrying about these issues," but I suspect you prefer honesty to feigned emotionality. Anyway, I have concluded that the *single best thing* we can do to improve education is to increase the assessment literacy of key constituencies such as educators, policymakers,

and parents of school-age children. This *single best thing* has the potential to pay off in counteracting much of the assessment dumbness in which we currently engage. I regard the acquisition of more widespread assessment literacy as *an absolutely necessary precursor* to greater educational improvement. Think about it—without accurate evidence regarding instructional impact, how can we ever know what works well? Finally, as suggested in Chapter 10, I believe we need to nurture the creation of a cadre of assessment consultants who can review the quality of various technical reports regarding educational testing, and then present truly comprehensible translations of those reports to assessment-literate audiences.

Evaluations with instructionally insensitive tests. When the AERA-APA-NCME joint *Standards* hit the streets in 2014, they established a far more clear demand for *purpose-supportive evidence* when evaluating an educational test's quality. This was a wonderful move by the architects of the joint *Standards,* for it made vivid the need to evaluate programs or educators on the basis of students' test results—but *only* when the test being used is accompanied by solid evidence indicating that the test is suitable for its designated evaluative mission. The regrettable truth is that most of today's educational tests are being used to evaluate schools and teachers without even a wisp of evidence supporting this evaluative use. *Instructional sensitivity* refers to the degree to which a test is capable of distinguishing between well-taught and poorly-taught students. Sadly, a great many serious educational mistakes are made because they use the results of tests to compare test-takers' performance, not differentiate among the instructional quality provided by teachers or schools. The sooner the world realizes the seriousness of this ubiquitous error, the sooner such mistakes can be corrected through the use of instructionally sensitive tests.

Formative assessment's underuse. Formative assessment works well, so well, indeed, that its underuse is quite heartrending. It needs to be employed much more widely for, as some

researchers have recently concluded, appropriate implementation of the formative-assessment process can literally *double* the speed of students' learning. Not to use such a powerful assessment-illuminated instructional process constitutes an educational sin against children. To secure the considerable learning gains available from the use of this potent classroom strategy, it is important for us to understand what formative assessment is—and what it is not. Then, the use of formative assessment must match different teachers' comfort in employing this process. Some teachers will be comfortable using formative assessment only occasionally; some teachers will prefer to use it constantly. Because we need to see more formative assessment in our nation's classroom, we should make certain that teachers do not become overwhelmed by formative assessment's required actions. If possible, in a formatively oriented classroom setting, students themselves should play a significant role in becoming self-directed and self-reinforcing learners. The formative-assessment process should, then, be vigorously promoted at every reasonable opportunity.

In *The ABCs* you'll find a Glossary and an Index. Please have fun with both. And thanks for your patience.