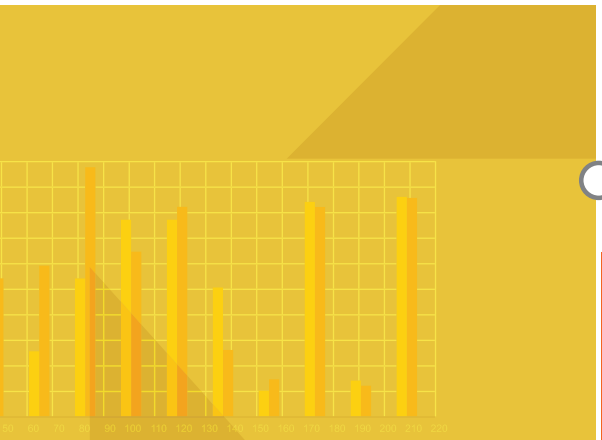
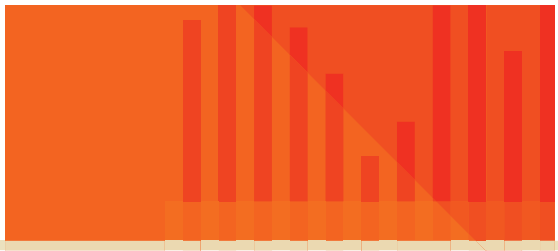
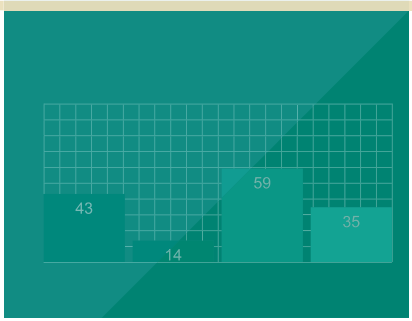




How to Make Decisions with Different Kinds of Student Assessment Data



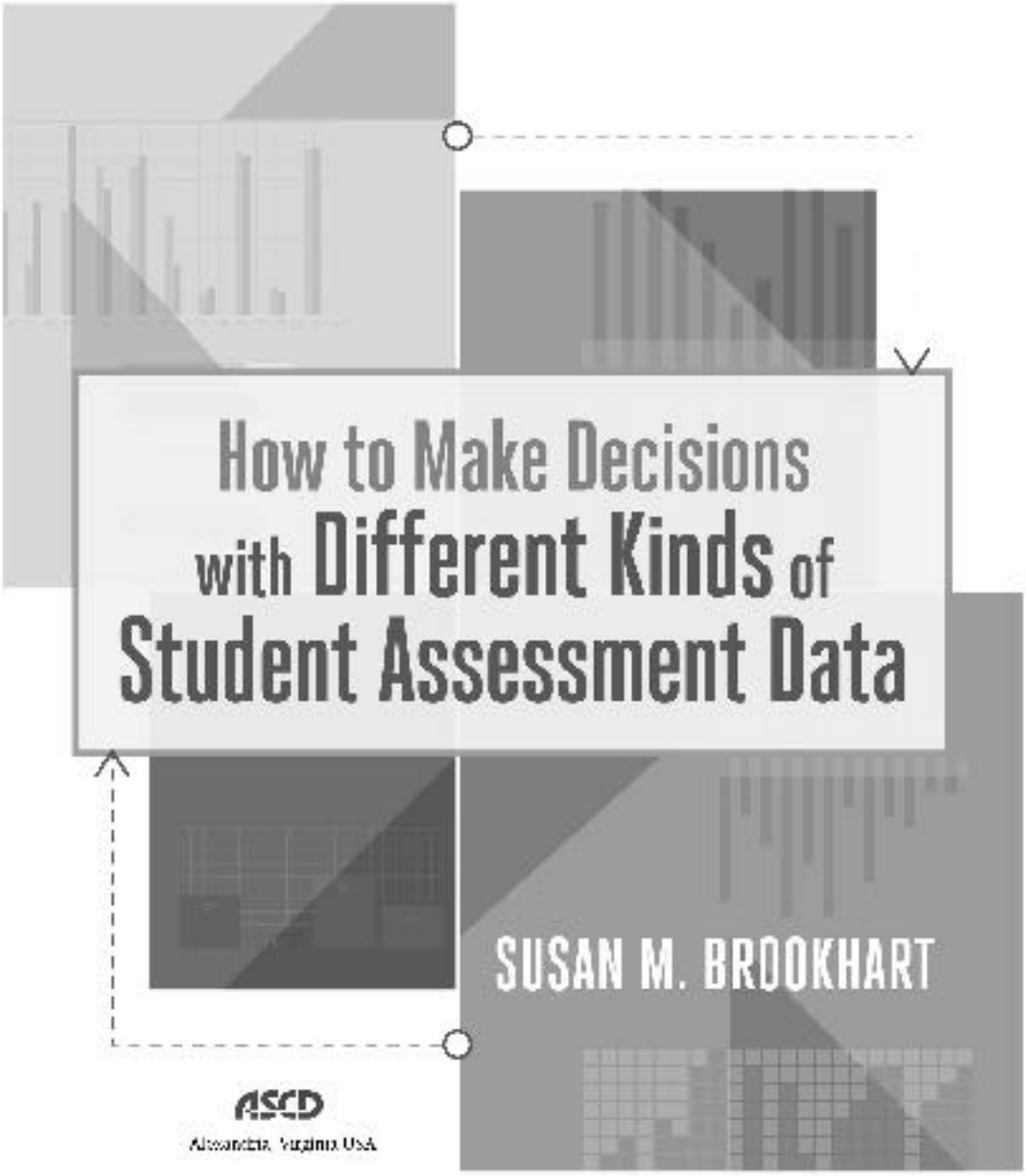
SUSAN M. BROOKHART

How to Make Decisions with Different Kinds of Student Assessment Data

ASCD MEMBER BOOK

Many ASCD members received this book as a member benefit upon its initial release.

Learn more at: **www.ascd.org/memberbooks**



1703 N. Beauregard St. • Alexandria, VA 22311 1714 USA
Phone: 800-933-2723 or 703-578-9600 • Fax: 703-575-5400
Website: www.ascd.org • E-mail: member@ascd.org
Author guidelines: www.ascd.org/write

Deborah S. Delisle, *Executive Director,* Stefani Roth, *Publisher;* Genny Ostertag, *Director, Content Acquisitions;* Julie Houtz, *Director, Book Editing & Production;* Darcie Russell, *Senior Associate Editor;* Georgia Park, *Senior Graphic Designer;* Mike Kalyan, Manager, Production Services; Valerie Younkin, *Production Designer; Typesetter*; Kelly Marshall, *Senior Production Specialist*

PAPERBACK ISBN: 978-1-4166-2103-4 ASCD product #116003
PDF E-BOOK ISBN: 978-1-4166-2105-8; see Books in Print for other formats.
Quantity discounts: 10–49, 10%; 50+, 15%; 1,000+, special discounts (e-mail programteam@ascd.org or call 800-933-2723, ext. 5773, or 703-575-5773). For desk copies, go to www.ascd.org/deskcopy.

ASCD Member Book No. FY15-3. ASCD Member Books mail to Premium (P), Select (S), and Institutional Plus (I+) members on this schedule: Jan, PSI+; Feb, P; Apr, PSI+; May, P; Jul, PSI+; Aug, P; Sep, PSI+; Nov, PSI+; Dec, P. For current details on membership, see www.ascd.org/membership.

Library of Congress Cataloging-in-Publication Data

[to be inserted]

How to Make Decisions with Different Kinds of Student Assessment Data

Acknowledgments vi

1 An Introduction to Different Kinds of Data 1

2 Large-Scale Accountability Assessments 18

3 Interim/Benchmark Assessments and Common Formative Assessments 46

4 Classroom Grades 63

5 Classroom Formative Assessment Strategies 78

6 Putting It All Together: Basing Decisions on Data 89

7 But Did They Learn Anything? Evaluating the Results of Your Decisions 105

8 Different Kinds of Data 121

Glossary 125

References 127

About the Author 132

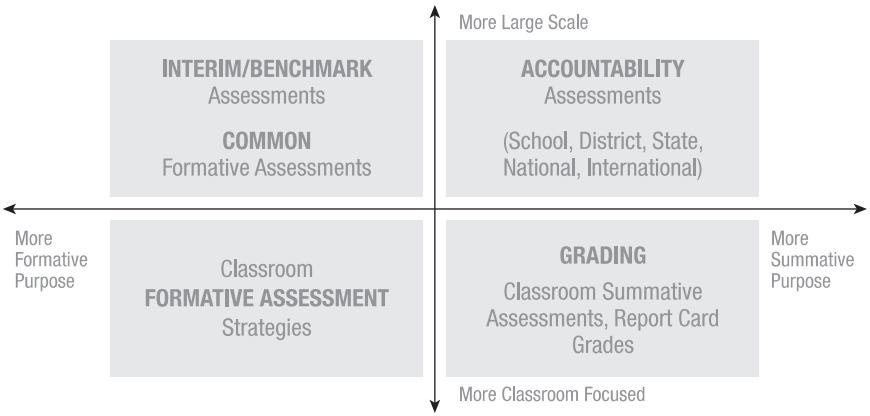
Index 133

Acknowledgments

The four-quadrant framework in this book has been in development for five years. I began the work at the invitation of Robert W. Lissitz, who invited me to present at the 2011 MARCES/MSDE (Maryland Assessment Research Center/Maryland State Department of Education) Conference *Informing the Practice of Teaching Using Formative and Interim Assessment: A Systems Approach* at the University of Maryland. I am grateful to him for posing the question of a systems approach, and I also thank the many people at that conference who encouraged me to develop the framework further. I am grateful to Margaret Heritage, who invited me to present this work at the Council of Chief State School Officers FAST SCASS (Formative Assessment for Students and Teachers State Collaborative on Assessment and Student Standards) in 2012, and to that organization for supporting further development of the framework. I am grateful to all the educators who have attended my assessment workshops and encouraged me to expand on the framework because they found it useful. This book is the result of their encouragement. I am grateful to my ASCD editors, Genny Ostertag and Darcie Russell: to Genny for believing this framework could be the basis for a book and to Darcie for editorial support. Of course, none of this work would have been possible without the love and support of my wonderful husband Frank and our daughters Carol and Rachel, to whom I am grateful for so much more than this book.

An Introduction to Different Kinds of Data

1



Simple logic does not always help us with data interpretation. Here are two examples, both of which are true stories.

Once during a workshop session, a school administrator explained to me that he saw standardized test scores as a sort of barometer. He aimed to make changes in his school that would lead to a rise in test scores, and that would be his indication that his reforms were successful. The analogy he gave me was the population of oysters in the Chesapeake Bay. Environmental reforms were needed in the bay area, changes were made, and the oyster numbers are increasing (see http://www.chesapeakebay.net/issues/issue/oysters for more information). Similarly, he explained, reforms in his school should lead to higher test scores.

The second example occurred at an airport, where I struck up a conversation with a businessman and his young son who were waiting for the same airport shuttle I was. The man said that he was really glad that his state now administered standardized tests, as per No Child Left Behind, because finally he had what he called a "bottom line" that he could watch to know how his child and his child's school were doing. His analogy was to the bottom line in a profit-and-loss statement in a business.

The administrator and the businessman dad were well-meaning people who valued education; they were not nay-sayers. They were both bright and successful individuals who applied logic and common sense to a problem they cared about. And they were both wrong.

Here's the thing. If you are an environmentalist or an oysterman, the oysters *are* the issue—or in the case of the Chesapeake Bay, one of the issues. Increasing the oyster population, to improve the habitat and the economy, *is* the purpose of the scientific reforms and management strategies. More oysters means the program is achieving its goal. Similarly, generating a profit *is* the purpose of being in business. Higher profits mean more money for shareholders, employees, and product development. Making money means your business is achieving its goal.

In contrast, raising test scores is not the purpose of education. The purpose of education has changed with society's needs and values over the years (Sloan, 2012). At this point in time, if you ask people the purpose of education, you will get answers such as these: to create adults who can compete in a global economy, to create informed citizens who can participate in the democratic process, to create critical thinkers and problem solvers, to create lifelong learners, or to create emotionally healthy adults who can engage in meaningful relationships. Obviously, no test score can tell you whether you have achieved these things.

The less obvious problem with our well-meaning administrator and businessman is that *even if you limit your interest to academic learning outcomes, raising test scores is not the purpose of education. The students' learning is.* Test scores are a measure of student learning, but they are not the thing itself. In our analogies, the oysters and the money were themselves the objects of interest. You can count oysters, and you can count money, but you can't "count" learning.

The best you can do to measure learning is to use a *mental measurement* that, if well designed, is a measure of learning in a limited domain. The key is to define clearly what that domain is, use a test or performance assessment that taps this domain in known ways, use a score scale with known properties that maps the student's performance back to the domain, and

interpret that score scale correctly when making inferences about student learning. The purpose of this book is to explain just enough about the properties of data on student learning so that you can make those inferences well. Then—and only then—can you make sound decisions. Another name for this purpose is *developing assessment literacy*. As the examples demonstrate, literacy in educational assessment involves more than counting or ranking. It involves specifying *what specific learning* you are measuring; understanding how the questions or tasks in the measure form a sample of that domain of learning; understanding properties of the scales, numbers, or categories used in the measurement; and being able to reason from all these things to make sound interpretations and decisions.

To complicate matters a bit, as the title of this book indicates, there are different kinds of data. For most educational decisions, you will want to mix the different kinds of data to broaden and deepen the pool of information about student learning that you use to monitor and improve that learning. You will want to know which kinds of data to watch, and when, in order to evaluate the effectiveness of your decisions. This book will help you do that in two ways. First, it offers a framework for thinking about assessment systems that categorizes different measures of student learning. Understanding how information differs from one category to another will help you interpret data. Second, this book offers some insights into different types of scores. Understanding different types of scores and their meanings will help you interpret data properly, as well. Equipped with an understanding of these two big ideas, your data interpretation and subsequent decisions will be more sound, more valid, and more useful.

The Purposes and Uses of Data

The phrase "data-based decision making" is used often and has many meanings. Teachers use data to answer questions about students. Groups of teachers and building administrators use data to answer questions about students, classes, programs, and their school. Central office administrators use data to make decisions about teachers, as well as students, classes, programs, and

schools. An Internet search on "data-based decision making" will bring up dozens of PowerPoint files, pdfs, images, and plans. Many books also address this theme.

It may seem like an obvious point, but the data you choose should be related to the intended purpose. If I want to make a tablecloth, I need to measure the length and width of the table; measuring its height won't help me much. The same principle operates in making decisions about student learning, but it's less easy to see. For example, if I want to make a decision about which reading skills to emphasize in my reading class, shouldn't I just look up students' scores on the state reading test? No. What the state reading test measures is general, overall "reading achievement," as defined by a whole set of reading standards. State test results will give you a sense of how your students do at "reading in general," as defined by whatever reading standards your state says its test measures, taken all together.

For example, the Smarter Balanced Assessment Consortium says that, regarding reading, its assessments can support this claim: "Students can read closely and analytically to comprehend a range of increasingly complex literary and informational texts" (Smarter Balanced, 2014). If I'm a teacher with a student whose assessment results suggest he can't do that very well, how do I design instruction for him? Low test results relative to this claim suggest a general decision—more or different reading instruction—but don't provide any clues about what aspects of reading to emphasize, remediate, or build on. To design reading instruction for that student, I'll need different data, assessment data that give a more fine-grained description of what the student can and cannot do as he reads.

In this simple example, reasoning from data is a two-step process. Data from the standardized accountability assessment help me identify a problem (Arland doesn't read proficiently) and lead me to another question: Why? To answer that second question, I need different data, because the reading accountability test doesn't give me information that is specific enough. Using complementary kinds of data for educational problem solving requires understanding different kinds of data.

This focus on a deeper understanding of data about student learning is what sets this book apart from other data books. I will, of course, also talk about how to use the data to inform instructional improvement. Two other excellent books that talk about using data to inform instructional improvement are *Data Wise* by Boudett, City, and Murnane (2013) and *Using Data to Focus Instructional Improvement* by James-Ward, Fisher, Frey, and Lapp (2013).

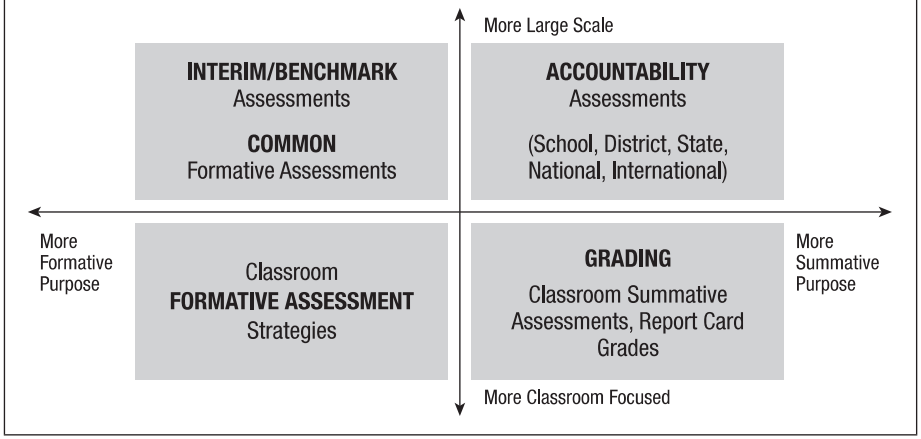
This book complements those and other books about data by focusing on developing a clearer understanding of exactly what test scores and other data about student learning *are* and what they *mean*. As an analogy, think about reading a Shakespeare play in a high school English class. If you read the play with a basic understanding of the English language, you will understand the plot. If you take the time to learn some Shakespearean vocabulary, you will understand the plot and the word-play nuances that unlock some of the humor and characterization in the play. In other words, you will understand the play better. Similarly, if you do data-based decision making with a basic understanding of assessment and of numbers, you will be able to make general decisions. If you learn some concepts about how educational assessments are constructed and some nuances about what their results mean, you will understand better what the data are telling you about your students' achievement.

Data About Student Learning

One way to organize and describe the different kinds of data about student learning is to use a four-quadrant framework (Brookhart, 2013). This framework allows us to group different kinds of data according to general type and purpose and to examine how they complement each other. It gives us some vocabulary to describe "assessment" in more specific terms. Figure 1.1 shows a four-quadrant framework for describing different kinds of assessments of student learning. The framework does not attempt to cover other data of interest to educators (e.g., student attendance, the number of books in the library, the ratio of students to teachers), just assessments of student learning.

FIGURE 1.1

Four-Quadrant Framework for Describing Types of Assessments



This framework will help you use different types of data to get richer, fuller information for your classroom decisions. The response of teachers and administrators to this framework has been quite positive. I have found that people are looking for a way to describe their "assessment system" that is more than just a long list of assessments.

Two dimensions: Purpose and focus

The framework in Figure 1.1 defines two dimensions on which assessment of student learning can be described: intended *purpose* for the information (formative or summative) and intended *focus* of the information (classroom or large scale). Of course, individual students are the ones who are assessed in all cases; even the large-scale state accountability test is administered to individual students. The focus dimension is about the place where the information is centered, and for large-scale assessments that focus is across individuals, classrooms, and schools.

Purpose. If learning is the main emphasis in education, then the distinction in purpose—between assessment that informs learning and

assessment that certifies that learning—is important. Many readers will be familiar with balancing formative and summative assessment in their classroom practice. Formative assessment, or assessment *for* learning, occurs during learning and is intended to result in improved learning (Moss & Brookhart, 2009). Summative assessment, or assessment *of* learning, occurs after an episode of learning and is intended to summarize the student's achievement level at a particular time (Moss, 2013). Typically, formative assessment items and tasks, and formative feedback, tackle next-step-sized learning targets. By the time summative assessment is appropriate, the outcome may be broader. As a somewhat oversimplified illustration, feedback on a 2nd grader's writing might be about capitalization and punctuation today and ideas tomorrow, and the final graded writing sample may appraise both.

Focus. The location of reference for the learning information is the other dimension—whether assessment is centered in the classroom or in a large-scale context. Some readers may be less familiar with distinguishing classroom-focused from large-scale assessments than they are with distinguishing formative and summative assessment purposes. After all, it's students in classrooms who take all the assessments, right? In Chapters 2 through 5, you will see that it is very useful to distinguish assessments that are primarily focused on the learning that occurs in one classroom, with its particular instructional context, from assessments that are primarily focused on generalizing across classroom contexts. The two differ in important ways, most notably on what specific learning is assessed and in the kinds of numbers that are used to quantify student performance. Some of that assessment information is meant for classroom use, and some is meant to be aggregated across classrooms for larger-scale evaluation—of a course, a program, or a school, for example.

Four quadrants

Crossing the two dimensions results in four quadrants that define the four major types of assessment of student learning that are used in schools, or what I have been calling "different kinds of data." These types are formative classroom assessment; interim/benchmark assessment, including "common

formative assessments" that are intended to be given in more than one classroom; summative (graded) classroom assessment; and summative (accountability) large-scale assessment. I'll briefly describe each type here and then devote a chapter to each.

Formative assessment: Formative purpose, classroom focus. Formative assessment is an active and intentional learning process that partners teachers and students to continuously and systematically gather evidence of learning with the express goal of improving student achievement (definition from Moss & Brookhart, 2009, p. 6; also see Wiliam, 2010). Formative assessment involves strategies such as the following (Moss & Brookhart, 2009; Wiliam, 2010):

- Sharing learning targets and criteria for success with students
- Feedback that feeds forward, from teachers, peers, or other sources
- Student self-assessment and goal setting
- Using strategic questions and engaging students in asking effective questions

One of the hallmarks of formative classroom assessment is student involvement. Formative assessment strategies aim to develop assessment-capable students who can see where they are headed (they can envision a learning target and know what it represents), take stock of where they are in relation to the target, and understand what they need to do next to continue to approach the learning target. Formative assessment's foundation is the students' clear concept of a learning target, or even a broader learning goal, and a clear understanding of what achievement of that goal looks like. This means students understand the criteria for success, or what Moss and I call "student look-fors" (Moss & Brookhart, 2009). Receiving feedback and using it to improve, setting goals and monitoring progress toward them, and asking effective questions all are based on the foundation of understanding "what I am trying to learn."

In recent years, people have quibbled over whether students have to be involved in making decisions about assessment results in order for the assessment to be formative. One of the reasons for the confusion is ignoring

the distinction between classroom-focused and large-scale assessment. Students have to be involved in *classroom* formative assessment, which works only when students take action on what they should do next in their learning. They can't take effective cognitive action if they aren't involved in making the decisions. However, students don't necessarily have to be involved in decisions made about large-scale assessments that are intended to be formative in purpose. Teachers may use these results to modify instruction, for example, without the students knowing about it.

I have found that the focus dimension—classroom versus large-scale—helps enormously with the vocabulary problem educators have been struggling with regarding formative assessment. For example, I was talking with a principal who thought "formative assessment" had to refer to the interim assessments his district purchased from a testing company, and so he didn't know what to make of the formative assessment strategies that occur within daily lessons—which was the topic I was at his school to address. I showed him the four-quadrant framework, and he found it immediately helpful. We might not be able to do much about the fact that the term *formative assessment* is currently used in too many different ways, but we certainly can make sure that we understand exactly what we are talking about for any specific data and interpret the data accordingly. This framework will help you do that.

Interim assessment: Formative purpose, large-scale focus. Teachers can use interim/benchmark assessments that do not involve students—other than to respond to assessment items or tasks—to inform instructional planning for those students or even for future students. This is a formative purpose, although it's not what I generally have in mind when I use the term *formative assessment*; to me, that term usually means classroom formative assessment. Interim assessment and classroom formative assessment are different from each other in a couple of ways. One difference is that interim assessment data can be aggregated, whereas classroom formative assessment data cannot. The second difference is that the users of interim assessment data are teachers, whereas the users of formative assessment data are students and teachers. The four-quadrant framework makes these differences explicit, because interim assessment is in the "large-scale" space, above the axis, and

formative classroom assessment is in the "classroom-focused" space, below the axis. In Chapters 3 and 5 I'll say more about what this means for interpreting and using assessment results.

There is a place for assessment information that can inform future instructional decisions and even administrative decisions. An example might be using interim assessments three times a year, each to inform planning of the next nine weeks of mathematics instruction, or combining information from several interim assessments to decide that, given limited funds, the school will hire a mathematics specialist next year instead of another classroom teacher. Some schools use assessments they call "common formative assessments," which they administer in more than one class, typically across a grade level and a subject area. If common formative assessments are used for planning (and not grading), they fit in this category.

Some interim/benchmark assessments or common formative assessments are also used for classroom grading or other evaluative purposes (Abrams & McMillan, 2013). This is not a recommended practice, and the four-quadrant framework helps us see why. Using data for several different purposes at once *is* possible, but it requires making sure that the data are valid for both purposes. Validating data for two different uses is very difficult to do, and the interaction of the different uses usually ends up changing the meaning of assessment results for both purposes (Koch & DeLuca, 2012). I'll have more to say about this in Chapter 5, as well.

Interim/benchmark assessment and common formative assessment involve practices such as the following:

- Using an item bank to construct tests at checkpoint times—for example, at the end of a unit or quarterly, for all students taking a certain subject in a certain grade
- Using teacher-made tests, and sometimes performance assessments, at checkpoint times—for example, at the end of a unit or quarterly, for all students taking a certain subject in a certain grade
- Using commercially published tests at checkpoint times
- Using curriculum-based progress-monitoring data

Grading: Summative purpose, classroom focus. Classroom grading comes in two forms: (1) individual grades—summative assessment via tests or performance assessments or any other graded assignments, and (2) report card grades (Brookhart, 2011). A report card grade is dependent on the quality of the information in the individual assignments on which it is based. Great procedures for report card summarizing cannot make up for poor-quality assessment information. On the other hand, high-quality assessment information, summarized poorly, does not produce a meaningful report card grade either. Both high-quality component information and summarizing methods that preserve intended meaning in the composite are necessary.

Classroom grades are a longstanding tradition in education and have been the subject of controversy for years (Brookhart, in press). Classroom grades typically report on unit-sized (or thereabouts) "chunks" of classroom learning. Classroom grades are typically expressed either as percentages or on short scales made up of achievement categories (e.g., *ABCDF* or Advanced, Proficient, Basic, Below Basic). Chapter 4 will discuss these points in more detail.

Large-scale accountability assessment: Summative purpose, large-scale focus. Finally, there is summative, large-scale accountability assessment. State tests fit here, as well as national and international comparison studies. Large-scale accountability assessment has been the subject of much interest and study recently (Perie, Park, & Klau, 2007), so I don't have to defend its place in the framework. Many educators realize the potential value of large-scale accountability tests but feel they may have more influence than they should (Shepard, 2000). Chapter 2 will describe how accountability tests typically furnish data about learning at a very large grain size (e.g., reading, mathematics) and explain how to interpret data that may be norm-referenced, criterion-referenced, or "standards-referenced," which draws a bit from both (Brookhart & Nitko, 2015).

Balanced Assessment

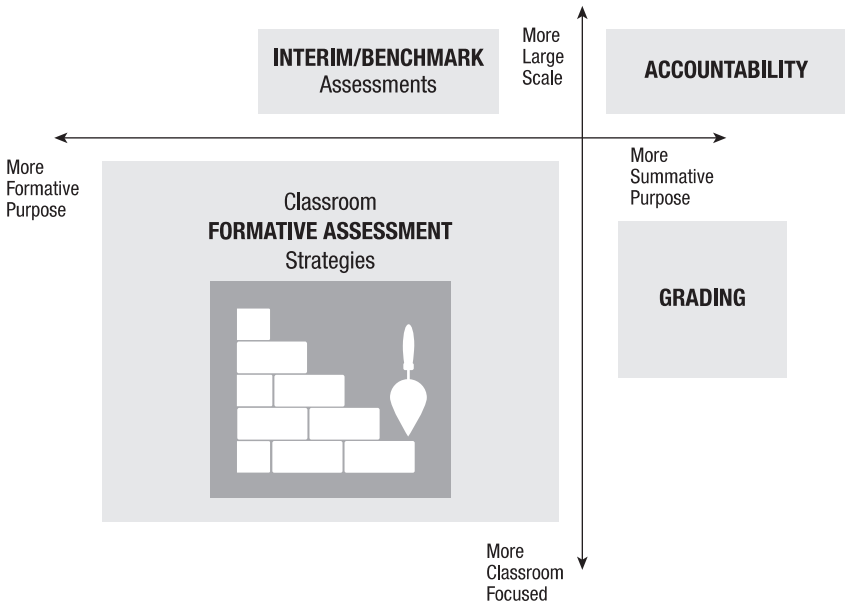
Looking at the framework, you might think that a balanced assessment system should simply contain equal amounts of assessment from of each of the four quadrants. This is not true.

Some materials designed for state and local administrators suggest that a balanced assessment system should be based on a tiered or pyramid-style model (e.g., Crane, 2010; Perie, Marion, & Gong, 2009). Classroom assessment, the bottom portion of the pyramid, is the most common or frequent kind of assessment, followed by interim assessment and then state accountability assessment. This pyramid model lumps all classroom assessment together and misses the opportunity to specify a good balance between classroom formative assessment and classroom summative assessment.

The four-quadrant framework privileges classroom-level data by devoting a dimension—and therefore half of the framework—to it. The classroom is the place where learning happens, and if we ignore the information closest to the learning, we lose a lot of fine-grained and diagnostic information. I argue for an assessment balance that looks more like Figure 1.2.

FIGURE 1.2

Achieving Balance in an Assessment System



Classroom formative assessment is the assessment that is best positioned to help improve learning. Well-designed classroom formative assessment focuses information on the growing edge of learning, puts the information in the hands of the students, and supports small, incremental, and immediate next steps in learning (Andrade, 2010; Black & Wiliam, 1998; Datnow & Hubbard, 2015; Moss & Brookhart, 2009). That's why Figure 1.2 enlarges the classroom formative assessment quadrant and suggests that this quadrant is the foundation of a balanced assessment system.

All types of assessment have a place, however. It is when data are used without an understanding of what they are *for* (e.g., a high school decides to use data from their benchmark assessment as part of students' report card grades to "make them take it seriously") that we get into trouble. I wrote this book to support informed use of the different kinds of data and to discourage using data for purposes for which the data are not suited.

How Sound Is Your Information?

The suitability of assessment results for particular purposes has a name: *validity*. This concept is central to educational assessment. Remember the businessman's bottom line? Because making money *is* a goal of his business, the amount of money made is one valid measure of the success of the business. Of course, it's not the only measure, because the business may have other goals, but clearly the bottom line is an important measure.

Contrast this with assessments that measure student learning. We can't directly measure that. We can't get "inside a kid's head." So we use "mental measures" that we construct according to sets of rules that we think might give us good estimates of student learning. Occasionally the set of rules is something like this: "Select a correct answer for each of these test items; for each answer you select, you will get one point; your score will be your total number of points divided by the number of possible points."

You can see there are several questions that need to be asked before we can be confident about what the score means. Some of those questions are about the test items. How much confidence you have that the score is a valid

measure of what you want to know depends on the answers to questions such as these:

- What were those test questions about? Were they about things the students had an opportunity to learn? Were they a representative mix of all the things the students were supposed to learn about the topic—that is, of all the possible questions you could have asked, how representative a sample is the set of questions you actually asked?
- Were the questions well written and clear? Were they at an appropriate level of difficulty? Did students understand all the words in the questions?
- Did students need any other specialized knowledge to answer the questions, besides what the questions were trying to test (e.g., did students need background knowledge on anything in a scenario, such as in a word problem or a reading passage) that might affect the test question's ability to measure achievement?
- Did the questions ask students to use thinking skills in the manner your learning outcomes intended (e.g., were students supposed to just recall information, or were they supposed to be able to use information to solve a problem or analyze a situation)?

Some of those questions are about the scoring procedure. How much confidence you have that the score is a valid measure of what you want to know also depends on the answers to questions such as these:

- Does each of those one-point questions really contribute about the same value or weight to the total score? Should some questions be worth more than others—and if so, why?
- Are there enough questions to ensure that the results provide you with an accurate sense of what students know and not just an indicator of chance performance? Are there enough questions to reasonably support using the percentage scale you have chosen (e.g., if there are only two questions, the only possible scores are 0, 50, and 100, which is a little silly).

This is not an exhaustive list of questions, but it should be enough to show you how you need to build a bridge of reasoning and evidence between

the score on any assessment and its meaning in a way that is not necessary for concrete measures like number of oysters or amount of money. This list of questions should also show you that the validity questions will differ for different kinds of data. Finally, the list should illustrate that behind questions about the meaning of a measure and its suitability for its purpose (validity), there are also some questions about the accuracy and consistency of the measure (reliability). You can't make meaningful decisions if you have inaccurate data.

As I describe the different kinds of data in Chapters 2 through 5, I'll try to highlight particular validity and reliability issues that are especially important for common uses of each kind of data. This book is not meant to be a substitute for a more thorough treatment of educational measurement (Brookhart & Nitko, 2015), but rather to be a guide to those who use data in practical ways in schools. To show that validity is important for everyone who uses educational data, not just for measurement experts, I ask you to consider two current examples that have validity questions at their core.

The issue of using large-scale, standardized tests as measures of school accountability with associated consequences has been with us for a while but became particularly salient when the No Child Left Behind Act was signed in 2002 (NCLB, 2002). Arguments about the suitability of large-scale tests to measure the quality of schooling range on both sides of the issue, and they hinge on validity questions. What do those tests measure? Are they the right indicators for the purpose of judging school quality? Are there other indicators that should be considered? Are the resulting judgments about school quality accurate?

Another current issue with validity questions at its core is the use of student assessment results as part of teacher evaluation systems (ISBE, 2013; NYSED, 2014). Arguments about the suitability of assessments for this purpose also range on both sides of the issue, and they too hinge on validity questions. What assessments are most appropriate for any given teacher, subject, or grade level? Is the student achievement thus measured really a measure of teacher quality? Do various statistical treatments applied to scores (e.g., pre-post gains, value-added models) appropriately remove the effects of irrelevant information and leave relevant information in the results? Are the resulting conclusions about teacher quality accurate?

Although these are hot-button issues, I believe that similar issues at the classroom level are equally important. Does your classroom assessment actually give you the information you think it does and support your next instructional move? Does your classroom assessment give your student an accurate picture of exactly what she needs to focus on next in order to improve in the way that she wants to? It always angers me that people say that daily classroom decisions are "less important" in some way than, say, medical decisions. Although it's true that poor medical decisions can lead to death, I always say that "messing with a kid's mind" is an equally dire consequence. I realize that an unintended wrong turn in instruction that might last five minutes and can be corrected in the next five minutes, but I submit that risks confusing the student or, at a minimum, wasting five minutes of precious learning time. We need to champion the cause of high-quality data for classroom use as well as large-scale uses, even though such data won't make headlines.

The Organization of This Book

Each of the next four chapters treats one quadrant of the framework. The chapter begins with a definition and a description of the kind or kinds of data available in that quadrant. I describe what you can (and can't) learn from these kinds of data, which in turn is based on the kind of learning and the grain size of the constructs that are commonly measured by that kind of data. Each chapter explains how to interpret common kinds of scores usually associated with that kind of data. Of course, I provide examples in each chapter.

In the process, I hope to demystify some quantitative concepts—not only different kinds of scores (e.g., what's the difference between a percentile rank and a percentage?), but also some principles (e.g., how aggregation can mask patterns in data, the difference between norm referencing and criterion referencing) and issues (e.g., the issues involved in measuring student growth, how to decide whether data are comparable—the "apples and oranges" issue). However, this is not a mathematics or statistics book. It is an "explaining" book. Just as you learned to read words, you can learn to "read" numbers.

Some words mean different things in different contexts. It may surprise you to learn the same is true for numbers.

The book concludes with three additional chapters. Chapter 6 discusses how to combine the different kinds of data about student learning, plus additional data about instruction, school culture, and resources, to answer questions about student learning and make decisions about what to do. Chapter 7 explains how to use different kinds of data about student learning to evaluate those decisions. Chapter 8 brings us back to the purpose for this book—to make you a better "reader" of data, who can reason with assessment results to arrive at defensible, effective decisions.

References

Abrams, L. M., & McMillan, J. H. (2013). The instructional influence of interim assessments: Voices from the field. In R. W. Lissitz (Ed.), *Informing the practice of teaching using formative and interim assessment* (pp. 105–133). Charlotte, NC: Information Age Publishing.

Andrade, H. L. (2010). Students as the definitive source of formative assessment: Academic self-assessment and self-regulation of learning. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 90–105). New York: Routledge.

Atkinson, R. C., & Geiser, S. (2009). Reflections on a century of college admissions tests. *Educational Researcher*, *38*(9), 665–676.

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher*, *36*(5), 258–267.

Bandeira de Mello, V., Blankenship, C., & McLaughlin, D. H. (2009). *Mapping state proficiency standards onto NAEP scales: 2005–2007* (NCES 2010-456). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Baum, M. H. (2011). *Using short-cycle interim assessment to improve educator evaluation, educator effectiveness, and student achievement.* Wisconsin Rapids, WI: Renaissance Learning.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, *5*(1), 7–74.

Black, P. J., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation, and Accountability*, *21*(1), 5–31.

Blanc, S., Christman, J. B., Liu, R., Mitchell, C., Travers, E., & Bulkley, K. (2010). Learning to learn from data: Benchmarks and instructional communities. *Peabody Journal of Education*, *85*(2), 205–225.

Boudett, K. P., City, E. A., & Murnane, R. J. (Eds.). (2013). *Data wise: A step-by-step guide to using assessment results to improve teaching and learning* (2nd ed.). Cambridge, MA: Harvard Education Press.

Bowers, A. J. (2010). Analyzing the longitudinal K–12 grading histories of entire cohorts of students: Grades, data driven decision making, dropping out and hierarchical cluster

analysis. *Practical Assessment, Research & Evaluation*, *15*(7). Available: http://pareonline.net/getvn.asp?v=15&n=7

Bowers, A. J., & Sprott, R. (2012). Examining the multiple trajectories associated with dropping out of high school: A growth mixture model analysis. *Journal of Educational Research*, *105*(3), 176–195.

Bowers, A. J., Sprott, R., & Taff, S. A. (2013). Do we know who will drop out? A review of the predictors of dropping out of high school: Precision, sensitivity and specificity. *The High School Journal*, *96*(2), 77–100.

Brookhart, S. M. (2001). Successful students' formative and summative uses of assessment information. *Assessment in Education*, *8*(2), 153–169.

Brookhart, S. M. (2007). Expanding views about formative classroom assessment: A review of the literature. In J. H. McMillan (Ed.), *Formative classroom assessment: Theory into practice* (pp. 43–62). New York: Teachers College Press.

Brookhart, S. M. (2011). *Grading and learning: Practices that support student achievement*. Bloomington, IN: Solution Tree.

Brookhart, S. M. (2013). Comprehensive assessment systems in service of learning: Getting the balance right. In R. W. Lissitz (Ed.), *Informing the practice of teaching using formative and interim assessment: A systems approach* (pp. 165–184). Charlotte, NC: Information Age Publishing.

Brookhart, S. M. (in press). Graded achievement, tested achievement, and validity. *Educational Assessment*.

Brookhart, S. M., Andolina, M., Zuza, M., & Furman, R. (2004). Minute math: An action research study of student self-assessment. *Educational Studies in Mathematics*, *57*(2), 213–227.

Brookhart, S. M., & Nitko, A. J. (2015). *Educational assessment of students* (7th ed.). Boston: Pearson.

Brown, R. S., & Coughlin, E. (2007, November). *The predictive validity of selected benchmark assessments used in the Mid-Atlantic Region* (Issues & Answers Report, REL 2007–No. 017). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic.

Castellano, K. E., & Ho, A. D. (2013, February). *A practitioner's guide to growth models*. Council of Chief State School Officers. Retrieved February 23, 2015, from http://www.ccsso.org/Resources/Publications/A_Practitioners_Guide_to_Growth_Models.html

Cizek, G. J. (Ed.). (2001). *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.

Crane, E. (2010). *Building an interim assessment system: A workbook for school districts*. Washington, DC: Council of Chief State School Officers.

Datnow, A., & Hubbard, L. (2015). Teachers' use of assessment data to inform instruction. *Teachers College Record, 117*(4), 1–26.

Gallimore, R., Ermeling, B. A., Saunders, W. M., & Goldenberg, C. (2009). Moving the learning of teaching closer to practice: Teacher education implications of school-based inquiry teams. *Elementary School Journal*, *109*(5), 537–553.

Goertz, M. E., Oláh, L. N., & Riggan, M. (2009). *From testing to teaching: The use of interim assessments in classroom instruction*. Consortium for Policy Research in Education Research Report # RR-65.

Hamilton, L., Halverson, R., Jackson, S. S., Mandinach, E., Supovitz, J. A., & Wayman, J. C. (2009). *Using student achievement data to support instructional decision making* (NCEE 2009-4067). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/wwc/PracticeGuide.aspx?sid=12

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112.

Heritage, M. (2010). *Formative assessment: Making it happen in the classroom*. Thousand Oaks, CA: Sage.

Ho, A. D. (2008). The problem with "proficiency": Limitations of statistics and policy under No Child Left Behind. *Educational Researcher*, *37*(6), 351–360.

Huff, K., Steinberg, L., & Matts, T. (2010). The promises and challenges of implementing evidence-centered design in large-scale assessment. *Applied Measurement in Education*, *23*(4), 310–324.

Illinois State Board of Education (ISBE). (2013, February). *Guidebook on student learning objectives for Type III assessments*. Illinois State Board of Education Performance Evaluation Advisory Council.

James-Ward, C., Fisher, D., Frey, N., & Lapp, D. (2013). *Using data to focus instructional improvement.* Alexandria, VA: ASCD.

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice*, *30*(4), 28–37.

Koch, M. J., & DeLuca, C. (2012). Rethinking validation in complex high-stakes assessment contexts. *Assessment in Education: Principles, Policy & Practice*, *19*(1), 99–116.

Konstantopoulos, S., Miller, S., van der Ploeg, A., Li, C.-H., & Traynor, A. (2011, September). *The impact of Indiana's system of diagnostic assessments on mathematics achievement*. Paper presented at the Fall 2011 Conference of the Society for Research on Educational Effectiveness, Washington, DC.

Leahy, S., Lyon, C., Thompson, M., & Wiliam, D. (2005, November). Classroom assessment: Minute by minute, day by day. *Educational Leadership*, *63*(3), 19–24.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education*, *6*(1), 83–102.

McIntosh, S. (2012, September). *State high school exit exams: A policy in transition*. Washington, DC: Center on Education Policy.

Miami-Dade County Public Schools. (2008, October). How interim assessments affect student achievement. *Information Capsule: Research Services*, vol. 0804. Available: http://drs.dadeschools.net/InformationCapsules/IC0804.PDF

Moss, C. M. (2013). Research on classroom summative assessment. In J. H. McMillan (Ed.), *Handbook of research on classroom assessment* (pp. 235–255). Los Angeles: Sage.

Moss, C. M., & Brookhart, S. M. (2009). *Advancing formative assessment in every classroom: A guide for instructional leaders*. Alexandria, VA: ASCD.

New York State Education Department (NYSED). (2014, March). *Guidance on the New York State district-wide growth goal-setting process for teachers: Student learning objectives*. Author.

No Child Left Behind Act of 2001. (2002). Pub. L. No. 107–110, 115 Stat. 1425.

Noffke, S. E., & Somekh, B. (Eds.). (2009). *The SAGE handbook of educational action research*. London: SAGE.

Northwest Evaluation Association (NWEA). (2012, January). *RIT Scale Norms Study: For use with Northwest Evaluation Association Measures of Academic Progress® (MAP®) and MAP for Primary Grades.* Portland, OR: Author.

PARCC. (2013, February). *PARCC college- and career-ready determination policy in English language arts/literacy and mathematics & policy-level performance level descriptors*. Adopted by the PARCC Governing Board and Advisory Committee on College Readiness, October 25, 2012, revised February 20, 2013. Retrieved February 23, 2015, from http://www.parcconline.org/CCRD

Pattison, E., Grodsky, E., & Muller, C. (2013). Is the sky falling? Grade inflation and the signaling power of grades. *Educational Researcher*, *42*(5), 259–265.

Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice*, *28*(3), 5–13.

Perie, M., Park, J., & Klau, K. (2007, December). *Key elements for educational accountability models*. Washington, DC: Council of Chief State School Officers.

Riggin, M., & Oláh, L. N. (2011). Locating interim assessments within teachers' assessment practice. *Educational Assessment*, *16*(1), 1–14.

Ross, J. A., Hogaboam-Gray, A., & Rolheiser, C. (2002). Student self-evaluation in grade 5–6 mathematics: Effects on problem-solving achievement. *Educational Assessment*, *8*(1), 43–58.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*(2), 119–144.

Saunders, W. M., Goldenberg, C. N., & Gallimore, R. (2009). Increasing achievement by focusing grade-level teams on improving classroom learning: A prospective, quasi-experimental study of Title 1 schools. *American Educational Research Journal*, *46*(4), 1006–1033.

Sawyer, R. (2013). Beyond correlations: Usefulness of high school GPA and test scores in making college admissions decisions. *Applied Measurement in Education, 26,* 89–112

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, *29*(7), 4–14.

Shepard, L. A. (2006). Classroom assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 623–646). Westport, CT: Praeger.

Sloan, W. M. (2012, July). What is the purpose of education? *Education Update*, *54*(7).

Smarter Balanced Assessment Consortium. (2013a, April). *Initial achievement level descriptors and college content-readiness policy: ELA*. Retrieved February 23, 2015, from http://www.smarterbalanced.org/achievement-levels/

Smarter Balanced Assessment Consortium. (2013b, April). *Initial achievement level descriptors and college content-readiness policy: Mathematics*. Retrieved February 23, 2015, from http://www.smarterbalanced.org/achievement-levels/

Smarter Balanced Assessment Consortium. (2014, February). *Appendix B: Grade level tables for all claims and assessment targets and item types*. Retrieved January 9, 2015, from http://www.smarterbalanced.org/wordpress/wp-content/uploads/2011/12/ELA-Literacy-Content-Specifications.pdf

Third International Conference on Assessment for Learning. (2009). Position paper on assessment for learning from the Third International Conference on Assessment for Learning, Dunedin, New Zealand. *Educational Measurement: Issues and Practice*, *28*(3), 3–4.

Thorsen, C., & Cliffordson, C. (2012). Teachers' grade assignment and the predictive validity of criterion-referenced grades. *Educational Research and Evaluation*, *18*(2), 153–172.

Waltman, K. K., & Frisbie, D. A. (1994). Parents' understanding of their children's report card grades. *Applied Measurement in Education*, *7*(3), 223–240.

Wiliam, D. (2010). An integrative summary of the research literature and implications for a new theory of formative assessment. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of Formative Assessment* (pp. 18–40). New York: Routledge.

Wiliam, D. (2011). *Embedded formative assessment*. Bloomington, IN: Solution Tree.

About the Author

Susan M. Brookhart, PhD, is an independent educational consultant based in Helena, Montana. She has taught both elementary and middle school. She was professor and chair of the Department of Educational Foundations and Leadership at Duquesne University, where she currently serves as an adjunct professor. She has been the education columnist for *National Forum*, the journal of Phi Kappa Phi, and editor of *Educational Measurement: Issues and Practice*, a journal of the National Council on Measurement in Education. She is the author or coauthor of several books, including ASCD's *How to Give Effective Feedback to Your Students a*nd *How to Create and Use Rubrics for Formative Assessment and Grading.* She is the coauthor, with Connie M. Moss, of ASCD's *Advancing Formative Assessment in Every Classroom: A Guide for Instructional Leaders, Formative Classroom Walkthroughs: Assessment in Every Classroom* and *Learning Targets: Helping Students Aim for Understanding in Today's Lesson*. She was named the 2014 Jason Millman Scholar by the Consortium for Research on Educational Effectiveness and Teaching Effectiveness (CREATE) and received the 2015 Samuel J. Messick Lecture Award from ETS/TOEFL. She may be reached at susanbrookhart@bresnan.net.

Related ASCD Resources: Student Assessment and Data

At the time of publication, the following ASCD resources were available (ASCD stock numbers appear in parentheses). For up-to-date information about ASCD resources, go to www.ascd.org.

ASCD EDge Group

Exchange ideas and connect with other educators interested in assessment on the social networking site ASCD EDge™ at http://ascdedge.ascd.org/

Print Products

Advancing Formative Assessment in Every Classroom: A Guide for Instructional Leaders by Connie M. Moss and Susan M. Brookhart (#109031)

Assessment and Student Success in a Differentiated Classroom by Carol Ann Tomlinson and Tonya R. Moon (#108028)

Checking for Understanding: Formative Assessment Techniques for Your Classroom, 2nd Edition by Douglas Fisher and Nancy Frey (#115011)

Formative Assessment Strategies for Every Classroom: An ASCD Action Tool, 2nd Edition by Susan M. Brookhart (#111005)

How Teachers Can Turn Data into Action by Daniel R. Venables (#114007)

Test Better, Teach Better: The Instructional Role of Assessment by W. James Popham (#102088)

Transformative Assessment by W. James Popham (#108018)

Transformative Assessment in Action: An Inside Look at Applying the Process by W. James Popham (#111008)

Using Data to Focus Instructional Improvement by Cheryl James-Ward, Douglas Fisher, Nancy Frey and Diane Lapp (#113003)

What Teachers Really Need to Know About Formative Assessment by Laura Greenstein (#110017)

The Whole Child Initiative helps schools and communities create learning environments that allow students to be healthy, safe, engaged, supported, and challenged. To learn more about other books and resources that relate to the whole child, visit www.wholechildeducation.org.

For more information: send e-mail to member@ascd.org; call 1-800-933-2723 or 703-578-9600, press 2; send a fax to 703-575-5400; or write to Information Services, ASCD, 1703 N. Beauregard St., Alexandria, VA 22311-1714 USA.