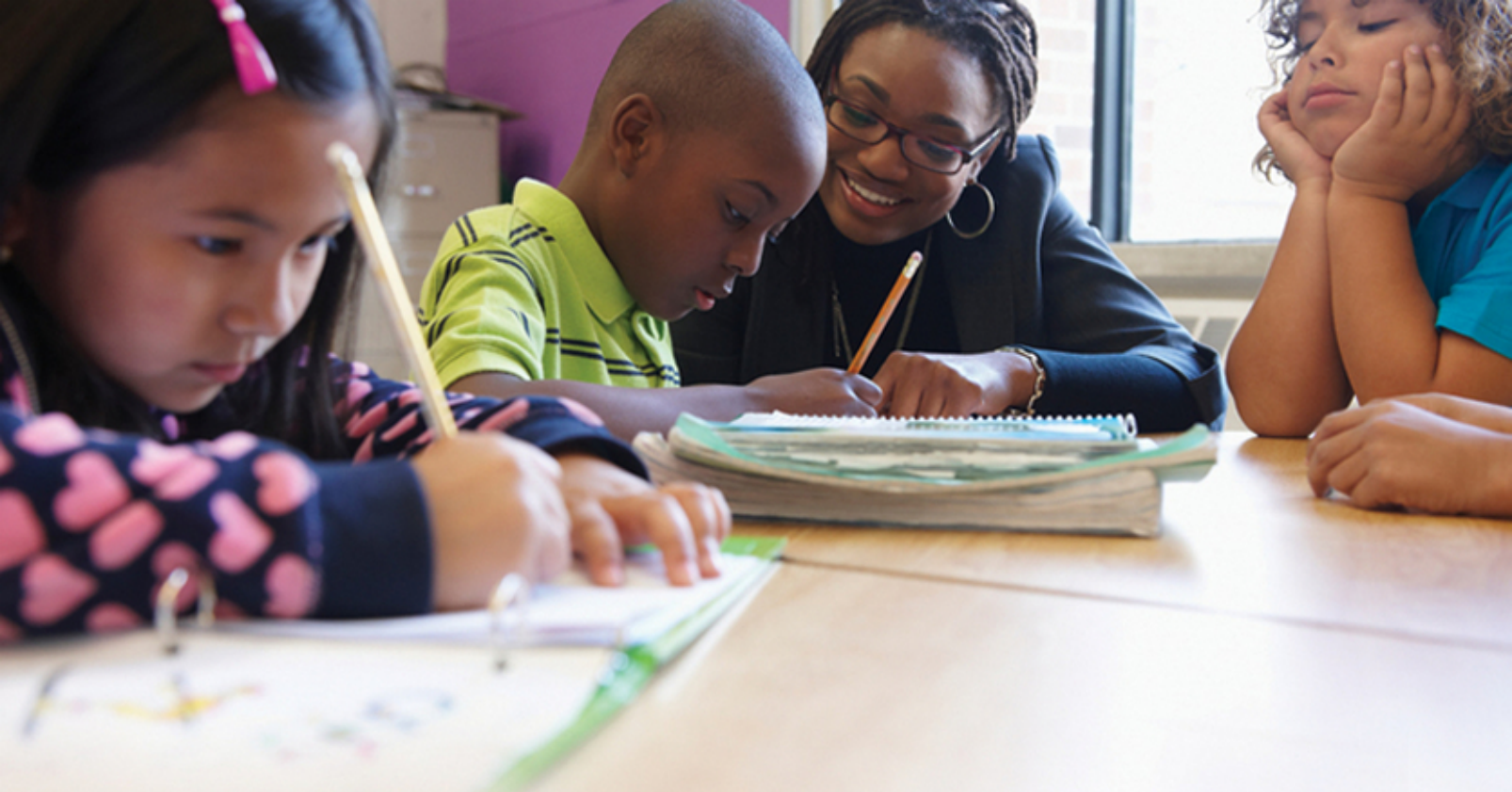

EDUCATIONAL TESTING & MEASUREMENT

CLASSROOM APPLICATION AND PRACTICE

• 10th Edition •

TOM KUBISZYN • GARY D. BORICH

TENTH EDITION

EDUCATIONAL TESTING AND MEASUREMENT:

Classroom Application and Practice

TENTH EDITION

EDUCATIONAL TESTING AND MEASUREMENT:
Classroom Application and Practice

TOM KUBISZYN
University of Houston

GARY BORICH
The University of Texas at Austin

WILEY

VICE PRESIDENT AND EXECUTIVE PUBLISHER	Jay O'Callaghan
SENIOR ACQUISITIONS EDITOR	Robert Johnston
ASSISTANT EDITOR	Brittany Cheetham
SENIOR CONTENT MANAGER	Lucille Buonocore
SENIOR PRODUCTION EDITOR	Anna Melhorn
SENIOR MARKETING MANAGER	Margaret Barrett
DESIGN DIRECTOR	Harry Nolan
PRODUCTION SERVICES	Suzanne Ingrao/Ingrao Associates
COVER PHOTO CREDIT	©LWA/Dann Tardif/Getty Images, Inc
COVER DESIGNER	Jasmine Lee

This book was set in 10/12 Times Roman by Laserwords and printed and bound by Donnelley Jefferson City. This book is printed on acid-free paper. ♾

Founded in 1807, John Wiley & Sons, Inc. has been a valued source of knowledge and understanding for more than 200 years, helping people around the world meet their needs and fulfill their aspirations. Our company is built on a foundation of principles that include responsibility to the communities we serve and where we live and work. In 2008, we launched a Corporate Citizenship Initiative, a global effort to address the environmental, social, economic, and ethical challenges we face in our business. Among the issues we are addressing are carbon impact, paper specifications and procurement, ethical conduct within our business and among our vendors, and community and charitable support. For more information, please visit our website: *www.wiley.com/go/citizenship*.

978-1-118-46649-0 (Main Book ISBN)
978-1-118-54005-3 (Binder Ready Version ISBN)

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

PREFACE

Welcome to the tenth edition of *Educational Testing and Measurement; Classroom Application and Practice*, an up-to-date, practical, reader-friendly resource that will help you navigate today's evolving and complex world of educational testing, assessment, and measurement. Users of the ninth edition of the text should have no difficulty recognizing and adapting to the tenth edition. Although Chapters 1, 2, 3, and 20 have been revised in substantive ways, these revisions inform readers about developments since the ninth edition but do not alter the sequencing either within these chapters or between the chapters. By comparison, revisions to the remaining chapters have been limited to clarifying language changes and the addition of references that reflect more recent knowledge and research.

The revisions to Chapter 1 underscore the importance of incorporating multiple sources of information about students whenever important educational decisions are made. We continue to advise against reliance on a single test administered at a single point in time to inform such decisions. We emphasize that this is especially true when important educational decisions must be made about the increasing numbers of students from diverse linguistic and cultural backgrounds in today's schools. In addition, we have updated recent findings from international testing programs that enable us to compare the performance of American students to those from both industrialized and nonindustrialized nations. We also expanded the section on competency testing for teachers to include a number of recent developments that will be of interest to most teachers in training.

The high-stakes testing (HST) chapter (Chapter 2) continues to inform readers about the history, issues, and controversies that surround HST, the differences between the No Child Left Behind (NCLB) Act and state HST programs, and offers recommendations for teachers that can help their students prepare for HST. Our revisions inform readers about several important developments that have occurred since the ninth edition of this text. These revisions include an explanation of the waivers that enable states to avoid penalties for failing to reach the NCLB goal of 100% proficiency in reading and math by the 2013–2014 school year, the development of the Common Core State Standards (CCSS) that have been adopted by all but five states at the time this revision went to press, and recent surveys suggesting that public support for HST may have diminished in the last few years.

Chapter 3 continues to inform readers about the elements of the Response to Intervention (RTI) approach to general and special education reform, the role of the regular classroom teacher in implementation of the RTI approach, and examples of how data collected from brief formative assessments informs data-based decision making. Because the RTI approach is new, our revisions to this chapter emphasize recently published research that addresses the technical and implementation challenges associated with this approach, which continues to be implemented unevenly across the country.

Chapter 20 continues to inform the reader about a variety of standardized achievement, academic aptitude, and personality assessment instruments. Our revisions are limited to the updating of those standardized tests that have been revised by their publishers

since we last described these measures in the ninth edition (TerraNova, Third Edition, CogAT Form 7; Minnesota Multiphasic Personality Inventory-2-RF, or MMPI-2-RF).

Since the ninth edition was published, the intensity of the controversy surrounding NCLB, high-stakes testing, and educational testing and assessment in general has not diminished. Of course, this makes it tempting to "take sides" and advocate for one position or another. However, as we have in all earlier editions of this text, our approach has been to present a balanced perspective, informed by the ever increasing research base. We have continued to strive to present both sides of the various controversies in hopes of enabling you to be informed enough to form your own opinions, and we continue this approach in this edition.

As with all previous editions of *Educational Testing and Measurement*, we continue to present complex test and measurement content in a friendly, nonintimidating, and unique manner, and to relate this content in meaningful ways to important developments in educational measurement and assessment. In completing this revision, we have kept our primary audience—classroom teachers—fully in mind. We have striven to present often abstract and sometimes difficult concepts and procedures in an up-to-date and accurate, but accessible, manner. Rather than overwhelm students with jargon and statistical theory, we continue to use a friendly, conversational style to enhance our emphasis on the application of theory. At the same time, we provide sufficient theoretical background to ensure that students will understand the foundations of measurement and avoid an oversimplified approach to measurement. Thus, we expect that both new and long-time users of the text should feel comfortable with the new edition of the text.

The chapter sequence remains the same as in the ninth edition. Two additional chapters devoted to testing and assessment of special learners, and the development of teacher-made instruments to assess student attitudes toward learning and student behavior, are available for review on the text's accompanying website (go to http://www.wiley.com/college/kubiszyn and click on the link to the Student Companion Site). The flexible organization of the text continues to enable instructors to either follow the chapter sequence as is or to modify it as needed to meet their particular needs.

As with earlier editions, readers will find at the conclusion of each chapter a step-by-step summary in which all important concepts in the chapter are identified for review, and a section of practice items and discussion questions. The discussion questions and exercises should help students learn how to apply the concepts presented, and, along with the Instructor's Manual (also available on the text's accompanying website) instructors will be able to identify organized, relevant activities and assignments that can be integrated into their class presentations. Discussion questions and exercises marked with an asterisk have answers listed in Appendix F.

We have tried to select traditional and contemporary topics and provide examples that help the teacher, especially the beginning teacher, deal with practical, day-to-day issues related to the testing and assessment of students and measuring their behavior, in the context of NCLB, state high-stakes testing programs, and RTI. The topics we have chosen, their natural sequences and linkage to the real-life tasks of teachers, the step-by-step summaries of major concepts, and our discussion questions and exercises, all work, we believe, to make this text a valuable tool and an important resource for observing, measuring, and understanding life in today's changing classroom. We hope that our approach helps ensure that these important activities are sensitive to the increasing accountability requirement today's educators face.

ACKNOWLEDGMENTS

We would like to express our appreciation to the following instructors for their constructive comments regarding this text over the years: Neal Schnoor, University of Nebraska–Kearney; Molly Jameson, Ball State University; Janet Carlson, University of Nebraska–Lincoln; Barry Morris, William Carey University; Lenore Kinne, Northern Kentucky University; Christopher Maglio, Truman State University; and Michael Trevisan, Washington State University. Thanks also to W. Robert Houston, University of Houston; Alice Corkill, University of Nevada—Las Vegas; Robert Paugh, University of Central Florida; Priscilla J. Hambrick, City University of New York; Pam Fernstrom, University of North Alabama; Bill Fisk, Clemson University; Lilia Ruban, University of Houston; David E. Tanner, California State University at Fresno; Gregory J. Cizek, University of Toledo; Thomas J. Sheeran, Niagara University; Jonathan A Plucker, Indiana University; Aimin Wang, Miami University; William M. Bechtol, late of Southwest Texas State University; Deborah E. Bennett, Purdue University; Jason Millman, Cornell University; David Payne, University of Georgia; Glen Nicholson, University of Arizona; Carol Mardell-Czudnowski, Northern Illinois University; and James Collins, University of Wyoming for their constructive comments on earlier revisions. Also, thanks to Marty Tombari for his contributions to Chapters 9 and 10 and other examples, illustrations, and test items in this volume, and to Ann Schulte for her contributions to Chapter 18. Finally, thank you to Aaron Boyce and Natalie Raff, doctoral students at the University of Houston, for their conscientious help in reviewing and refining the tenth edition.

—Tom Kubiszyn and Gary Borich

CONTENTS

CHAPTER 1

AN INTRODUCTION TO CONTEMPORARY EDUCATIONAL TESTING AND MEASUREMENT

CHANCES ARE that some of your strongest childhood and adolescent memories include taking tests in school. More recently, you probably remember taking a great number of tests in college. If your experiences are like those of most of the students who come through our educational system, you probably have very strong or mixed feelings about tests and testing. Indeed, some of you may swear that you will never test your students when you become teachers, unless of course you are required by law to do so! If so, you may think that test results add little to the educational process and fail to reflect learning, that testing may turn off students, or that tests do not measure what they are supposed to measure. Others may believe that tests are necessary and vital to the educational process. For you, they may represent irrefutable evidence that learning has occurred. Rather than viewing tests as deterrents that turn off students, you may see them as motivators that stimulate students to study and provide them with feedback about their achievement.

Between those who feel positively about tests and those who feel negatively about them lies a third group. Within this group, which includes the authors, are those who see tests as tools that can make important contributions to the process of evaluating pupils, curricula, and teaching methods, but who question the status and power often given to individual tests and test scores. We are concerned that test users and consumers of test results (e.g., teachers, parents, the media, administrators, policy makers, and other decision makers) often uncritically accept test scores without considering how useful the test scores may actually be for whatever decision may be at hand.

TESTS ARE ONLY TOOLS; THEIR USEFULNESS CAN VARY

Uncritical acceptance of test scores by decision makers concerns us for five reasons. First, tests are only tools, and tools can be appropriately used, unintentionally misused, and intentionally abused. Second, tests, like other tools, can be well designed or poorly

designed. Third, both poorly designed tools and well-designed tools in the hands of ill-trained or inexperienced users can be dangerous. Fourth, the usefulness of a well-designed tool, even in the hands of a competent user, can be limited if the tool, or test, is used for an unintended purpose or population. In other words, just as there is no "one-size-fits-all" tool (not even the venerable Swiss army knife!), no single test is appropriate for all purposes and all persons. Fifth, even when a test is well designed and is appropriately used by a competent examiner (i.e., for the purpose and populations it was designed for), the test can only provide us with *some* of the information we may want or need to make the best possible educational decision about a student.

"Wait a minute!" you may say. "All this makes it sound like you're saying that tests are not useful for educational decision making, even if they are well constructed and properly used." Not so! We are *not* saying test results are not useful, unimportant, or unhelpful. We *are* saying that it *is important* to recognize that the usefulness of tests, like the usefulness of all tools, depends on a variety of factors. Let's explore some of these factors next.

WHY WE DEVELOPED THIS TEXT: ENHANCING TEST USEFULNESS

The five concerns we mentioned above helped motivate us to write this text. By helping you learn to design and to use tests and test results appropriately, we hope you will be less likely to misuse tests and their results and be better able to recognize and avoid using poorly designed tests. We also hope that you will become mindful of how the purpose of testing and the population to be tested can affect a test's usefulness. Finally, we hope that you will grasp the importance of considering multiple sources of information obtained from multiple informants *along with* test results to make important educational decisions about students. Let's turn to a more detailed explanation of how each of these points can affect the usefulness of a test for educational decision making.

TECHNICAL ADEQUACY

A critically important factor that affects a test's usefulness is its technical adequacy. Much of this text is devoted to helping you develop teacher-constructed (or teacher-made) tests with good technical adequacy and in helping you evaluate the technical adequacy of commercial tests (i.e., developed by test publishers). The technical adequacy of a test includes evidence of its validity (see Chapter 15) and its score reliability (see Chapter 16). Validity evidence helps us determine whether the test is measuring what it is intended to measure, and score reliability indicates the extent to which test scores are consistent and stable. In general, we strive to use tests with the strongest validity and score reliability evidence. However, these factors are *not fixed* characteristics of a test, even if the test is well established, widely used, and respected. This is because a test's validity and score reliability can be affected by many factors, including the competency of the test user, whether the test is being used for the purpose it was developed, the person or population it is used with, and even the testing conditions (e.g., noisy rooms, poor lighting, timing errors) (see Chapters 15–19).

This is why we said before that no test is a "one-size-fits-all" test that is equally useful for all test users, purposes, and populations. Thus it is inappropriate to speak of the "validity of a test" or the "reliability of a test," as though validity and reliability are permanent, unchanging characteristics of the test. Nevertheless, this is exactly what many test users believe. Because test usefulness can vary, it is most appropriate to speak of the evidence of a test's validity and score reliability for a particular use and with a particular population, when administered by a competent test user. The need to require test user competency and to clarify a test's intended use and the intended population when discussing the test's usefulness emerged from deliberations among measurement experts from the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) over several years when they developed the latest edition of *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999). The *Standards*, which are currently undergoing revision, are widely regarded as one of the most authoritative and influential guidelines in the testing, measurement, and assessment arena. Next, let's consider several examples that should help you understand why it is so important to consider the competency of a test user and a test's validity and reliability evidence within the context of a test's intended use and the intended population.

Test User Competency

Evidence of a test's usefulness can vary depending on the competency of the people administering, scoring, and interpreting the test. An electric drill (the corded type, not the battery-powered ones that always seem to need recharging) can be very useful in the hands of a competent electrician who is skilled in carefully drilling holes in a wall while avoiding the electrical and water lines behind the wall. The same drill may be far less useful, and even dangerous, in the hands of a child or in the hands of an adult who acts like a child! Does this mean that a child or an incompetent adult could not drill a hole in the wall? Of course not; it simply means that the competent tool user will make better use of the tool, just as a competent test user will likely make better use of a test. Could a child use the drill to drill holes in the wall? Probably. Could the child avoid all the electrical and water lines behind the wall? And could the child avoid drilling a hole in his hand or electrocuting himself? We can only hope! In short, an electric drill's, or a test's usefulness, varies depending on the competency of the person using it.

Matching the Test's Intended Purpose

A screwdriver is intended to be used to drive screws. Nonetheless, who hasn't used a screwdriver as an ice pick, a lever or pry bar, a chisel, a paint mixing stick, a means to poke an older sibling in the eye, or for some other purpose? Did it work? Probably. Did it work as well as an ice pick, a lever or pry bar, a chisel, or a sharp stick would have? Probably not. In short, the screwdriver's usefulness depends on whether you are using it for its intended purpose.

Specific Purposes Like other tools, tests have been designed for many specific measurement purposes (achievement in various academic content areas, intellectual and personality functioning, vocational aptitudes, etc.). Like other tools, a test's usefulness (i.e., the evidence of its validity and reliability) can vary, depending on how well the

current purpose of testing matches the specific purpose for which the test was developed. A test designed to identify individuals with above-average ability to quickly and accurately recognize typographical errors in a document may have excellent validity and score reliability when it is used to predict a potential employee's ability to quickly and accurately recognize typographical errors in a book manuscript. On the other hand, the validity of the same test may be substantially lower if the test is used to predict a person's ability to actually write a book (a very different skill, believe us!). In this case, the test's usefulness is more limited. This does not mean it is useless for this purpose, but there may be better tools . . . oops, we mean tests, that would be more useful.

General Purposes: Formative and Summative Assessment In addition to being designed for a wide variety of specific content areas (e.g., assessing reading vocabulary, comprehension, spelling, mathematics, algebra, general science), educational tests also can be designed for the more general purposes of *formative* and *summative* assessment. Summative tests/assessments have been and continue to be the most commonly administered tests in education (see Chapters 2 and 19 for examples). Summative tests are administered after some period of instruction (this can vary widely, e.g., a unit on vertebrates in biology, a semester of physics, a year of algebra) and are intended to provide a measure or gauge of student learning following the *completion* of a unit of instruction. Summative tests are lengthy and are used to assign grades, evaluate curriculum effectiveness, assess annual gains in student, school, and district academic improvement (i.e., to meet state and federal accountability requirements), and for a variety of other purposes. Summative tests/assessments can be very useful if the purpose of testing is to inform us about broad achievement trends *after* instruction has been completed. However, summative tests/assessments may not be very useful if the purpose of testing is to evaluate the effectiveness of instruction on a day-to-day basis. Summative tests are simply not designed to be sensitive to such small, specific changes in achievement; rather, they are designed to measure larger and broader changes in achievement.

Formative tests/assessments will be more useful than summative assessments if the purpose of testing is to inform day-to-day instructional decision making (e.g., move on to the next step in the curriculum, review or re-present the content using a different approach/medium, or provide instruction in a different setting). Formative assessments tend to be brief so as to minimize interference with instructional time and to facilitate repeated administration in the classroom. One type of formative assessment is called *curriculum-based measurement*, or CBM (Jones, 2008; Hosp, Hosp, & Howell, 2006). CBM assessments are called probes, and these probes are about one minute long. CBM probes are intended to be utilized on an ongoing, frequent basis as part of the instructional process to monitor student progress (i.e., progress monitoring).

The frequent administration of such brief, formative tests/assessments enables the teacher to make daily adjustments to instruction, as necessary, to maximize student learning. When frequent progress monitoring indicates a student is not progressing at the same rate as are other students in the class, this may indicate the student needs either more intensive or differently delivered instruction. Formative assessments such as CBM have played a relatively minor role in classroom testing in the past. However, the passage of the *Individuals with Disabilities Education Improvement Act* (IDEIA) in 2004 prompted a rapid and dramatic increase in the use of formative assessment for progress monitoring of student learning in the *regular* education classroom over the last

few years (Federal Education Budget Project, 2011). If you know that the IDEIA is a *special* education law, you may wonder how it could affect regular education in this way.

The answer is that today's classrooms have been transformed by many diverse populations into heterogeneous classrooms of culturally, linguistically, and academically diverse learners. This change has brought many of the concerns historically relegated to special education front and center into the regular classroom. We will elaborate on this very recent phenomenon and its significant and growing impact on the regular classroom teacher later in this chapter and in more detail in Chapter 3.

Matching Diverse Test-Takers to the Test

As you no doubt know, our population has become increasingly diverse in recent years, and there is no reason to expect that this trend will diminish any time soon. This trend is reflected in today's increasingly diverse classrooms, where a wide range of cultural, linguistic, and academic backgrounds are common (Banks & Banks, 2009). Yet, the technical adequacy of many educational tests and assessments was established based on samples that included primarily, if not entirely, Caucasian, Hispanic American, and African American students. Would we expect the technical adequacy of these tests to be the same when used with populations from different cultural, linguistic and academic backgrounds (e.g., Middle Eastern and Indonesian learners, limited English-speaking learners, and higher and lower socioeconomic learners)? Before you answer this question, let's return to the example of the electric drill.

Did you ever try to drill a hole in metal with a drill bit designed for drilling into wood? If you did, you will not make that mistake again! Specialized drill bits have been developed to enhance usefulness when drilling into diverse surfaces (e.g., wood, metal, concrete, ceramic). Thus, a wood bit works best for drilling into wood, a metal bit for drilling into metal, and so on. Would we expect that one bit would work equally well for all diverse surfaces? Of course not. To be most useful, the drill bit must match the surface into which you're drilling the hole.

Things are no different with tests. For example, a test may be designed to measure certain characteristics for a particular group. We would expect the test to be most useful when used for similar groups, but like a single drill bit, we would not expect the test to be equally useful for all groups. Let's consider a multiple-choice test used to assess end-of-year achievement in a tenth-grade American History class (quick, is this a formative or a summative test?). The class is composed largely of two groups: native English-speaking U.S. children who have taken many multiple-choice tests and recent immigrants from Nicaragua who speak little English and have had little formal schooling. Excellent evidence may exist for the test's technical adequacy (e.g., reliability and validity) when it is used to assess achievement for the English-speaking students. However, evidence for the same test's reliability and validity may be less impressive (or even nonexistent) when used to assess achievement for the limited English proficient (LEP), recently emigrated students from Nicaragua. In this case, linguistic (lack of familiarity with the English language) and cultural (lack of experience with multiple-choice tests in their native Nicaragua) factors may seriously limit the usefulness of the test.

This consideration does not necessarily mean the test should not be used at all with this population. It does mean we should always be thoughtful and try to select the test that is most useful for the population we are testing—if there is one available—and to be very careful in interpreting results.

Although using a test that is not a great "fit" is not ideal, it is a matter of practicality. We simply lack tests that have strong evidence for their validity and usefulness with all populations and for all purposes for which tests are used. For example, school children in the Houston, Texas, public schools speak almost 200 different languages and have a similarly wide range of cultural backgrounds. Because of the diverse cultural, linguistic, and academic backgrounds of these students, it follows that the usefulness of the tests used to evaluate these diverse children will vary. This leads us to our next point.

Although we should always strive to select the test that is *most* useful for the group(s) to be tested, we cannot always achieve this goal. When we cannot match the purpose and the group, we should try to be *especially* thoughtful and careful in interpreting test results. What else can we do?

Test Results and Diversity Considerations

What should you do when the group being tested does not match the characteristics of the sample used in its development? Depending on whom you ask, you will get a variety of suggestions. Here are ours. In such situations, the results of a single test administered at a single point in time should *never* be used alone to make important decisions—even when the technical adequacy, test user competency, and purpose criteria we have just described have been met. Instead, we recommend that testing should be part of a thoughtful, multifaceted approach to assessment, with input provided over time by multiple informants (i.e., teachers and other trained personnel). In our diverse society there can be no "one-size-fits-all" test or assessment.

That, as they say, is the theory (or perhaps wishful thinking on our parts). Reality is very different. Promotion, graduation, and other high-stakes educational decisions (e.g., ranking of schools as exemplary, acceptable, or in need of improvement) are commonly made based entirely, or primarily, on test scores obtained at a single point in time, in spite of the increasingly diverse nature of our society. This phenomenon is largely attributable to the rapid spread of the high-stakes testing movement since the mid-1990s (which we will discuss in detail in Chapter 2).

That said, efforts have been undertaken to make accommodations for culturally, linguistically, and academically diverse test-takers (Flanagan, Ortiz & Alfonso, 2007). In some cases, tests have been translated or otherwise modified in an effort to better align them with diverse populations (Malda, van de Vijver, & Temane, 2010). Nevertheless, the technical adequacy of these modifications, together with their fairness to test-takers, have proven difficult to determine, and more study is clearly needed. Fairbairn and Fox (2009) provide a summary of the relevant issues, and they also offer test development suggestions for English-language learners.

TESTS ARE ONLY TOOLS: A VIDEO BEATS A PHOTO

The importance of making decisions based on more than a single test result is not a concern limited to testing with diverse populations. Even when a test has technical adequacy, the test user is competent, and the purpose and population are appropriate, we *still* do not recommend making important educational decisions based on a single test administered at a single point in time. Instead of relying on such a limited "snapshot"

or photograph (or JPEG) of student achievement for important decision making, we recommend that test results should be considered to be part of a broader "video" or process of measurement called assessment. We will describe the process of assessment in the next section and also distinguish between testing and assessment. See Box 1-1 about the Waco, Texas, public schools for an example of the controversial use of test results from a single test at a single point in time to make important educational decisions.

BOX 1-1

WACO, TEXAS, SCHOOLS USE STANDARDIZED TEST SCORES ALONE TO MAKE PROMOTION DECISIONS

Concerned with possible negative effects of social promotion, the Waco, Texas, public schools decided to utilize standardized test scores as the basis for promotion decisions beginning with first graders in 1998. As a result, the number of students retained increased from 2% in 1997 to 20% in 1998 (The Waco Experiment, 1998). The Waco schools are not alone in curtailing social promotion. The Chicago public schools, in the midst of a wide-ranging series of educational reform initiatives, retained 22,000 students in 1994, with 175,000 retained in 1998 (*Newsweek*, June 22, 1998).

Social promotion is a practice that purports to protect student self-esteem by promoting students to the next grade so that they may stay with their classmates even when they are not academically ready for promotion. Educational, psychological, political, fiscal, cultural, and other controversies are all associated with social promotion. What has come to be known by some as the "Waco Experiment" also raised a number of measurement-related issues.

Although the Waco schools' decision was doubtless well intended, their policy may have overlooked the fact that the utility of test scores varies with age, with test results for young children being less stable and more prone to error than those for older children. A relatively poor score on a test may disappear in a few days, weeks, or months after additional development has occurred, regardless of achievement. In addition, older children are less susceptible to distractions and, with years of test-taking experience under their belts, are less likely to be confused by the tests or to have difficulty completing tests properly. All these factors can negatively affect a student's score and result in a score that underrepresents the student's true level of knowledge.

Furthermore, a single standardized test score provides only a portion of a child's achievement over the school year, regardless of the grade level. As we will see when we consider the interpretation of standardized test results in Chapter 19, a number of student-related factors (e.g., illness, emotional upset) and administrative factors (e.g., allowing too little time, failing to read instructions verbatim) can negatively affect a student's performance on the day the test was taken. Thus, making a decision that so substantially affects a child's education based on a single measure obtained on a single day rather than relying on a compilation of measures (tests, ratings, observations, grades on assessments and portfolios, homework, etc.) obtained over the course of the school year seems ill-advised.

On the other hand, using data collected on a single day and from a single test to make what otherwise would be complex, time-consuming, and difficult decisions has obvious attraction. It appears to be expedient, accurate, and cost-effective and to be addressing concerns about the social promotion issue. However, it also may be simplistic and shortsighted if no plan exists to remediate those who are retained. As noted in a June 12, 1998, editorial in the *Austin American-Statesman*, "Failing students who don't meet a minimum average score, without a good plan to help them improve, is the fast track to calamity." Nevertheless, this trend has not diminished since we first reported on it in our sixth edition. Indeed, reliance on the use of test scores to make high-stakes promotion decisions has increased across the nation. Several states have now adopted versions of Florida's retention policy, enacted by then Governor Jeb Bush in 2002–2003 to combat social promotion. In these states, students who do not pass the states' high-stakes test must be retained, although there are often several "good cause" exemptions from this policy that soften this practice (Robelen, 2012).

So, unfortunately, the situation described in Box 1-1 is not unusual. Well-intended educators continue to rely solely or primarily on test results from a single point in time to make important, high-stakes educational decisions. At times, they may have little choice because federal, state, or district requirements mandate "one-size-fits-all" policies that are tied to scores from a specific, "approved" test, without regard for the extent to which the validity and reliability of the scores from this test may vary for diverse populations of students, or for a different purpose than the one for which the test was developed.

To sum up, our position is that tests are only tools that can be appropriately used, abused, or misused. To minimize inappropriate test use, it is important to carefully consider the (a) evidence of a test's technical adequacy, (b) competency of the test users, (c) extent to which the purpose of testing matches the purpose for which the test was developed, and (d) degree to which the test-takers match the group that was used to establish the technical adequacy of the test. Furthermore, we encourage you to consider additional background, historical, and observational data, especially when the test is administered to a group that differs from the test's development sample, and when the test is used to make high-stakes educational decisions (Rhodes, Ochoa, & Ortiz, 2005). In short, these situations call for an assessment process rather than simply testing/assessment.

DEFINING SOME TEST-RELATED TERMS

So far, we have clarified the notion that tests are only tools, and we have described some of the factors that can affect the usefulness of these tools. Next, we need to clarify some technical test-related terminology. The terms we introduce will be referred to over and over again in the text. Although it is important to understand as many of these terms as you can at this point, if you're like most students, you will need to return to this section repeatedly as you work your way through the text.

Tests, Assessments, and the Assessment Process

Today, the terms *tests* and *assessments* are commonly used interchangeably. Indeed, some seem to have eliminated the word "testing" from their vocabularies and replaced it with the word "assessment" because they believe that use of the word "assessment" is less evaluative, threatening, or negative than use of the word "testing." In any case, we too will consider the terms *testing* and *assessment* to be synonymous. However, we believe a clear distinction needs to be made between tests and assessments and the *assessment process*.

Tests and Assessments The terms *tests* and *assessments* typically refer to single measures that yield results at a single point in time. There are exceptions, and some of these (i.e., performance and portfolio assessments) will be discussed in Chapters 9 and 10. It is from the results of tests and assessments that we attempt to measure learning or to quantify some attribute or characteristic (e.g., intellectual ability, level of anxiety). Educational tests/assessments may be either formative or summative, depending on whether they are used to measure day-to-day changes in learning (i.e., formative) or learning over a more extended time frame (i.e., summative).

Assessment Process The assessment *process*, on the other hand, may span days, weeks, an entire semester, the entire school year, or longer. *Both* formative and summative assessments are typically part of this broad assessment process. The assessment process is a comprehensive evaluation made up of many testing and assessment components and relevant background and contextual information. A comprehensive assessment process may include the following:

a. Traditional (i.e., summative) test results from one or more multiple-choice, true-false, matching, or essay tests.

b. Progress monitoring (i.e., formative) results from less traditional tests such as curriculum-based measurement or CBM probes (to be described later in this chapter and in Chapter 3).

c. A variety of other measurement procedures (e.g., performance and portfolio assessments, covered later in the text, and observations, checklists, rating scales—included in the supplemental chapters on the textbook website at http://www.wiley.com/college/kubiszyn).

d. The findings from all these assessments are integrated with relevant background and contextual information (e.g., language proficiency and cultural considerations—also covered later in the text) to help ensure that educational decisions are appropriate and as valid as possible.

So, you can see that from our perspective, testing is only one part (i.e., like a snapshot or photograph) of the *process* of assessment that may include multiple photographs or segments (i.e., like a slide show, movie, or video) that reflect multiple types of information obtained from multiple informants at multiple points in time. Taken together, these components can provide us with a far richer and, we believe, more valid and accurate description of the individual than we can possibly obtain from any of the individual components alone. Figure 1.1 further clarifies the distinction between testing/assessment,

Testing/Assessment

1. Tests (or assessments) are developed or selected, administered to the class, and scored.
2. Test results may then be used to make decisions about a pupil (assign a grade, recommend for an advanced program), instruction (repeat, review, move on), curriculum (replace, revise), or other educational factors.

An Assessment Process

1. Information is collected from tests and *other measurement instruments* (portfolios and performance assessments, rating scales, checklists, and observations).
2. This information is critically evaluated and integrated with relevant background and contextual information.
3. The integration of critically analyzed test results and other information results in a decision about a pupil (assign a grade, recommend for an advanced program), instruction (repeat, review, move on), curriculum (replace, revise), or other educational factors.

FIGURE 1.1 The distinction between testing/assessment and the assessment process.