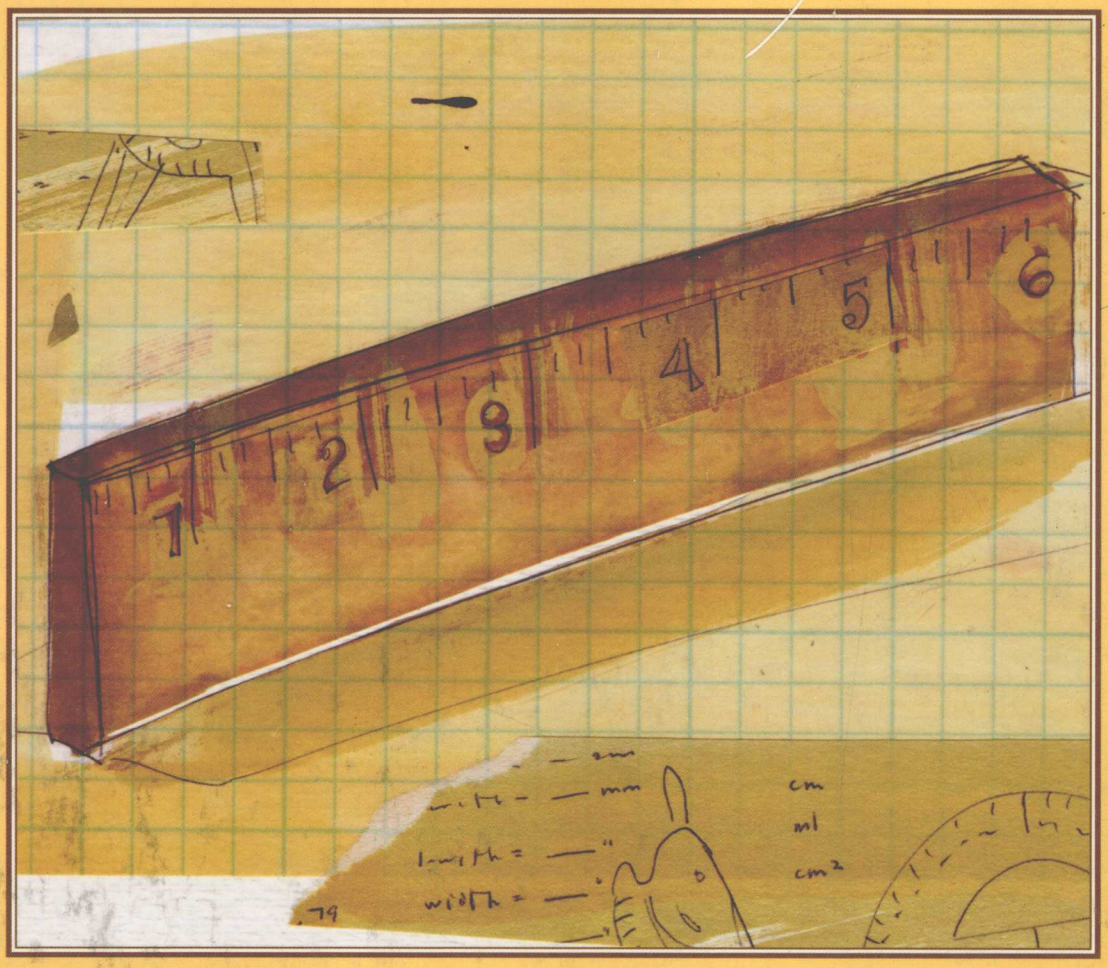
TOM KUBISZYN
GARY BORICH

Educational Testing and Measurement

Classroom Application and Practice



Seventh Edition

SEVENTH EDITION

EDUCATIONAL TESTING AND MEASUREMENT

Classroom Application and Practice

TOM KUBISZYN
The University of Texas at Austin

GARY BORICH
The University of Texas at Austin

JOHN WILEY & SONS, INC.

Acquisitions Editor *Brad Hanson*
Marketing Manager *Kevin Molloy*
Senior Production Editor *Valerie A. Vargas*
Senior Designer *Harry Nolan*
Production Management Services *Argosy*

This book was set in 10/12 Times Roman by Argosy and printed and bound by R. R. Donnelley & Sons Company. The cover was printed by Phoenix Color Corporation.

This book is printed on acid-free paper. 

Library of Congress Cataloging in Publication Data:
Kubiszyn, Tom.
Educational testing and measurement : classroom application and practice / Tom Kubiszyn, Gary Borich.—7th ed.
p. cm.
Includes bibliographical references and index.

1. Educational tests and measurements--United States. I. Borich, Gary D. II. Title.

LB3051 .K8 2003 371.26'0973--dc21 2002026733
ISBN 0-471-14977-2 (cloth : alk. paper)

Printed in the United States of America

10 9 8 7 6 5 4 3 2

EDUCATIONAL TESTING AND MEASUREMENT

Classroom Application and Practice

PREFACE

Two major developments in classroom testing and measurement explain why we have decided to incorporate substantial additions and revisions to this, the seventh edition of *Educational Testing and Measurement.* These developments were the rapid spread of high-stakes testing to all 50 states and the District of Columbia in the past few years and growing awareness among regular education teachers about their increased responsibility for special education students under the 1997 amendments to the Individuals with Disabilities Act (IDEA–97).

High-stakes test scores are now widely used for student promotion and graduation decisions and for educational accountability purposes, sometimes with substantial school or district incentives and penalties tied to student performance on high-stakes tests. And, with the passage in January 2002 of the "No Child Left Behind Act," we now have a federal mandate that will soon require annual academic assessments of all school children in grades 3–8. This ensures continued and probably increased attention to high-stakes tests in the foreseeable future. Under the IDEA–97 regular education teachers are now required to play a much broader role than in the past in the instruction and assessment of special education students included in regular education classrooms.

Because these developments have generated intense controversy (i.e., especially the rapid spread of high-stakes testing), one of the goals of this revision was to inform instructors and future teachers about these important developments in a balanced and thoughtful way. And, because all future teachers will have to cope with the demands of high-stakes testing and full compliance with IDEA–97, another goal was to provide future teachers with practical information and recommendations they can immediately use in the classroom to prepare themselves and their students for high-stakes testing and the challenges of IDEA–97. Nevertheless, as important as these developments are, the overarching goal of this revision was to remain true to the friendly style, content, order of presentation, and length of past editions of *Educational Testing and Measurement.*

As with all previous editions we have continued 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 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, long-time users of the text should continue to feel comfortable with it.

Past users of the text should have no difficulty recognizing and adapting to this revision. The overall organization has been only slightly modified, and the flexible organization of the text continues to enable instructors to either follow the chapter sequence as is or modify it as needed to meet their particular needs. A new chapter has been added (Chapter 2, High-Stakes Testing), another has been significantly expanded (Chapter 7, Writing Essay and Higher Order Test Items), and several other chapters have been revised and updated to seamlessly integrate the new material on high-stakes testing and IDEA–97 and other developments. To help keep the text's length reasonable the section on planning a schoolwide testing program has been deleted from Chapter 19 since this function has become obsolete in the face of the adoption by all states of state-mandated high-stakes tests. Other changes to the seventh edition are described in more detail next.

Chapter 1 has been revised and updated. It continues to provide up-to-date information on the increasingly important distinction between testing and assessment and new information about a variety of contemporary trends, especially high-stakes testing, the implications of IDEA–97 for regular education teachers, and competency testing for teachers.

Chapter 2 is a new chapter devoted to the high-stakes testing phenomenon. It defines high-stakes testing, traces its history, reviews both sides of the controversy surrounding the use of high-stakes tests, considers the position taken by national measurement associations, and provides future teachers with concrete recommendations they can use to prepare themselves and their students for high-stakes tests.

Chapter 3, which was Chapter 2 in previous editions, has been updated.

Chapter 4, Norm- and Criterion-Referenced Tests and Content Validity Evidence, consolidates Chapters 3 and 4 from previous editions into a single chapter. Several reviewers suggested combining Chapters 3 and 4 into a single chapter because of their brevity. To minimize confusion we have maintained the same topic sequence as in previous editions. Throughout Chapter 4 and in several later chapters we have substituted "validity evidence" for "validity" when appropriate to ensure continuity with language included in the most recent edition of the *Standards for Educational and Psychological Tests* (American Educational Research Association, 1999).

Chapters 5 and 6 also have been updated.

Chapter 7, Writing Essay and Higher Order Test Items, has been substantially revised and expanded. It now includes a wider variety of examples of essay items to help teachers see how they can be used to measure higher order thinking and problem-solving ability. The sections on scoring also have been revised and updated. And, a new section has been added to help teachers assess how well students can organize and access knowledge and another new section provides guidance and many examples to help teachers design and utilize open-book questions and tests.

Chapter 8, Performance-Based Assessment, was Chapter 9 in previous editions.

Chapter 9, Portfolio Assessment, was Chapter 10 in previous editions.

Chapter 10, Administering, Analyzing, and Improving the Test, was Chapter 8 in previous editions.

These chapters were reordered at the recommendation of reviewers who noted that many principles covered in Administering, Analyzing, and Improving the Test applied to performance and portfolio assessments, and not just to objective and essay items. In previous editions this chapter followed the chapters on objective and essay items but preceded the

chapters on performance and portfolio assessments. In this edition this chapter now follows all four chapters devoted to classroom-based assessment-objective items, essay and higher order test items, performance assessments, and portfolio assessments.

Chapters 11–14 have been updated.

Chapter 15, Validity, has been revised to make it consistent with the approach to the establishment of validity evidence described in the latest edition of the *Standards for Educational and Psychological Tests* (American Educational Research Association, 1999). Rather than considering validity to be a characteristic of a test, the new edition stresses the importance of acquiring evidence of a test's validity for a particular use.

Chapters 16 and 17 have been updated.

Chapter 18, Standardized Tests, has been revised. It continues its extensive treatment of the history, utility, and interpretation of standardized tests, with increased attention paid to the use of both standardized norm-referenced and standardized criterion-referenced tests in high-stakes testing programs.

Chapter 19, Types of Standardized Tests, has been revised and information regarding various standardized tests has been updated. The entire section entitled "Planning a School- or District-wide Testing Program" has been deleted because this function has been supplanted by state legislatures and state education agencies with the spread of the high-stakes testing phenomenon.

Chapters 20 and 21 have been revised and updated to better inform regular education teachers about their increased responsibilities for evaluating the educational and behavioral progress of special education students included in their regular education classrooms and curricula under IDEA–97. At the request of reviewers, person-first language (e.g., children with disabilities) has replaced the language previously used to refer to children in special education programs (i.e., special learners). New examples of recently developed or revised behavior rating scales that regular education teachers are increasingly expected to complete are included in Chapter 21. These include scales used to assess medication safety and efficacy for the growing number of pupils taking medications that can affect learning and behavior.

Finally, Chapter 22 also has been revised to reflect the rapid spread of the high-stakes testing phenomenon and the added responsibilities for regular education teachers for compliance with IDEA–97.

Throughout the text we have added references to a variety of contemporary measurement trends, tying these to day-to-day decision making for the classroom teacher. And, we have updated our references, suggested readings, and list of supplemental statistics and measurement texts to include recent articles, chapters, and books that reinforce and expand the changing face of educational measurement in today's classroom.

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. Additionally, we have prepared new discussion questions and/or exercises for each new chapter and section. These discussion questions and exercises should help students learn how to apply the concepts presented and, along with the *Instructor's Manual,* should help instructors identify organized activities and assignments that can be integrated into their class presentations. Discussion questions and exercises marked with an asterisk have answers listed in Appendix D.

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. 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.

Tom Kubiszyn
Gary Borich

ACKNOWLEDGMENTS

We would like to express our appreciation to the following instructors for their constructive comments on this revision: 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; and Pam Fernstrom, University of North Alabama. Thanks also are owed to Bill Fisk, Clemson University; 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, we thank Marty Tombari for his contributions to Chapters 8 and 9 and other examples, illustrations, and test items in this volume and Ann Schulte for her contributions to Chapter 17. Finally, we thank Denise Branley for her revisions to the *Instructor's Manual* and Test Bank.

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 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. If so, you may think that test results add little to the educational process and fail to reflect learning or that testing may turn off students to learning. 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 view 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.

TESTS ARE ONLY TOOLS

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 contribute importantly to the process of evaluating pupils, the curriculum, and teaching methods, but who question the status and power often given to tests and test scores. We are concerned that test users often uncritically accept test scores. This concerns us for three 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. These three concerns motivated 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.

TESTS ARE NOT INFALLIBLE

Test misuse and abuse can occur when users of test results are unaware of the factors that can influence the usefulness of test scores. The technical adequacy of a test, or its validity (see Chapter 15) and reliability (see Chapter 16), is one such factor. A variety of factors can dramatically affect the validity and reliability of a test. When a test's validity and reliability are impaired, test results should be interpreted very cautiously, if at all. Too often, such considerations are overlooked or ignored by professionals and casual observers alike.

Even when a test is technically adequate, misuse and abuse can occur because technical adequacy does not ensure that test scores are accurate or meaningful (see Chapters 17 and 18). A number of factors can affect the accuracy and meaningfulness of test scores. These include the test's appropriateness for the purpose of testing, the test's content validity evidence (if it is an achievement test), the appropriateness of its norms table (if it is a norm-referenced test, a term we will learn more about in Chapter 3), the appropriateness of the reading level, the language proficiency and cultural characteristics of the students, teacher and pupil factors that may have affected administration procedures and scoring of the test, and the pupils' motivation and engagement with the test on the test day.

Because technical adequacy and these interpretive factors can affect test scores dramatically, our position is that test scores should never be uncritically employed as the sole basis for important educational decision making. Nevertheless, with the rapid spread of the high-stakes testing movement (which we will discuss in detail in Chapter 2), a disturbing number of promotion and graduation decisions are being based on test scores alone. Instead of relying on such a limited "snapshot" of student achievement for important decision making, we recommend that test results should be considered to be part of a broader "movie" or process called *assessment*. It should be the findings of the broad assessment, not just test results, that form the basis for important educational decision making. We will describe the process of assessment in the next section and distinguish between testing and assessment. See the sidebar about the Waco, Texas, public schools for a recent example of the controversial use of test results alone to make important educational decisions.

TESTING: PART OF ASSESSMENT

Unfortunately, the situation described in the sidebar is not unusual. Well-intended educators continue to rely solely or primarily on test results to make important educational decisions. They may unintentionally misuse test results because they have come to regard test results as the end point rather than an early or midpoint in the much broader process of assessment. Or, they may mistakenly believe that testing and assessment are synonymous.

In the assessment process, test results are subject to critical study according to established measurement principles. If important educational decisions are to be made, critically evaluated test results should be combined with results from a variety of other measurement procedures (e.g., performance and portfolio assessments, observations, checklists, rating scales—all covered later in the text), as appropriate, and integrated with relevant background and contextual information (e.g., reading level, language proficiency, cultural considerations—also covered later in the text) to ensure that the educational decisions are

BOX 1-1

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

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 students are not academically ready for promotion.

Educational, psychological, political, fiscal, cultural, and other controversies are all associated with social promotion.

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 (*Austin American-Statesman,* June 12, 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).

What has come to be known by some as the "Waco Experiment" also raised a number of measurement related issues.

Whereas the Waco schools' decision was doubtless well intended, their policy may have overlooked the fact that the utility of test scores varies dependent on age, with test results for young children 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, irrespective 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 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 18, there are 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) that 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 (i.e., 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 appears 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, the use of test scores to make high-stakes promotion decisions has increased across the nation. We will explore this phenomenon in depth in Chapter 2.

appropriate. You can see that although testing is one part of assessment, assessment encompasses much more than testing. Figure 1.1 further clarifies the distinction between testing and assessment.

Throughout the text we will refer to testing and/or assessment. To avoid confusion later, note the distinction between testing and assessment in Figure 1.1. Next, we will summarize why we believe it is of vital importance that all educators obtain a firm grounding in educational testing and assessment practice.