EDUCATIONAL TESTING AND MEASUREMENT:
Classroom Application and Practice

TOM KUBISZYN
University of Houston

GARY BORICH
The University of Texas at Austin
# CONTENTS

## CHAPTER 1  AN INTRODUCTION TO CONTEMPORARY EDUCATIONAL TESTING AND MEASUREMENT  1

Tests Are Only Tools  1  
Tests Are Not Infallible: Usefulness Can Vary Across  
  Purposes And Persons  2  
  Different Purposes: Effects on Test Usefulness  2  
  Different Populations: Effects on Test Usefulness  3  
Testing: Only a Part of the Assessment Process  4  
Recent History of Educational Measurement  6  
  The No Child Left Behind Act (NCLB)  8  
  The Rapid Spread of State High-Stakes Testing Mandates  11  
2004 Reauthorization of the Individuals with Disabilities Education Improvement Act (IDEA–04)  12  
The Need to Identify Pupils at-Risk for Low Performance on Accountability Measures  13  
The Lake Wobegon Effect and High-Stakes Testing  13  
Will We See a National Test?  14  
Performance and Portfolio Assessment  15  
Education Reform and the Global Economy  16  
Competency Testing for Teachers  17  
Increased Interest from Professional Groups  18  
A Professional Association–Book Publisher Information Initiative  18  
Effects on the Classroom Teacher  20  
About The Text  22  
What if You’re “No Good in Math”?  23  
Summary  23  
For Discussion  25  

## CHAPTER 2  HIGH-STAKES TESTING  26

Comparing and Contrasting NCLB and State High-Stakes Testing Programs  26  
High-Stakes Testing: A Nationwide Phenomenon  27  
High-Stakes Tests Are Only Tools  29  
What Is High-Stakes Testing and Why Does It Matter?  30  
Promotion and Graduation Decisions Affect Students  30  
Principal and Teacher Incentives Are Linked to Performance  32  
Effects on Property Values, Business Decisions, and Politics  32  
The History of High-Stakes Testing  33  
  Education Reform  33  
  Standards-Based Reform  33  
The High-Stakes Testing Backlash  41  
Is There Really a High-Stakes Testing Backlash?  42  
What Do National Organizations Say about High-Stakes Tests?  44  
  AERA’s Twelve Conditions That High-Stakes Testing Programs Should Meet  45  
  How Can a Teacher Use the Twelve Conditions?  47  
Helping Students (And Yourself) Prepare For High-Stakes Tests  48  
  Focus on the Task, Not Your Feelings about It  48  
  Inform Students and Parents about the Importance of the Test  49  
  Teach Test-Taking Skills as Part of Regular Instruction  50  
  As the Test Day Approaches Respond to Student Questions Openly and Directly  51  
  Take Advantage of Whatever Preparation Materials Are Available  52  
Summary  52  
For Discussion  53  

## CHAPTER 3  THE PURPOSE OF TESTING  54

Testing, Accountability, and the Classroom Teacher  55  
  Types of Educational Decisions  56  
  A Pinch of Salt  60  
  “Pinching” in the Classroom  61  
What To Measure  62  
How To Measure  63  
  Written Tests  63  
Summary  65  
For Discussion  65
**CHAPTER 4  NORM-AND CRITERION-REFERENCED TESTS AND CONTENT VALIDITY EVIDENCE** 66

Defining Norm-Referenced and Criterion-Referenced Tests 66
Comparing Norm-Referenced and Criterion-Referenced Tests 70
Differences In the Construction of Norm-Referenced and Criterion-Referenced Tests 71
Norm- and Criterion-Referenced Tests and Linguistic, Cultural, and Contextual Sensitivity 72
Norm- and Criterion-Referenced Tests and Validity Evidence 74
A Three-Stage Model of Classroom Measurement 75
Why Objectives? Why Not Just Write Test Items? 77
Where Do Goals Come From? 79
Are There Different Kinds of Goals and Objectives? 80
How Can Instructional Objectives Make a Teacher’s Job Easier? 82
Summary 84
For Discussion 85

**CHAPTER 5  MEASURING LEARNING OUTCOMES** 87

Writing Instructional Objectives 87
Identifying Learning Outcomes 87
Identifying Observable and Directly Measurable Learning Outcomes 88
Stating Conditions 89
Stating Criterion Levels 90
Keeping It Simple and Straightforward 91
Matching Test Items To Instructional Objectives 92
Taxonomy of Educational Objectives 94
Cognitive Domain 94
Affective Domain 97
The Psychomotor Domain 99
The Test Blueprint 100
Content Outline 102
Categories 102
Number of Items 103
Functions 103
Summary 105
For Practice 105

**CHAPTER 6  WRITING OBJECTIVE TEST ITEMS** 107

Which Format? 107
True-False Items 109
Suggestions for Writing True-False Items 112
Matching Items 112
Suggestions for Writing Matching Items 115
Multiple-Choice Items 116
Higher Level Multiple-Choice Questions 122
Suggestions for Writing Multiple-Choice Items 125
Completion Items 126
Suggestions for Writing Completion or Supply Items 128
Gender and Racial Bias In Test Items 129
Guidelines for Writing Test Items 129
Advantages and Disadvantages of Different Objective Item Formats 131
Summary 132
For practice 133

**CHAPTER 7  WRITING ESSAY AND HIGHER ORDER TEST ITEMS** 134

What Is an Essay Item? 135
Essay Items Should Measure Complex Cognitive Skills or Processes 135
Essay Items: Extended or Restricted Response 136
Examples of Restricted Response Essays 138
Pros and Cons of Essay Items 140
Advantages of the Essay Item 140
Disadvantages of the Essay Item 140
Suggestions for Writing Essay Items 141
Scoring Essay Questions 143
Scoring Extended Response and Higher Level Questions 144
General Essay Scoring Suggestions 149
Assessing Knowledge Organization 150
Open-Book Questions and Exams 152
Some Open-Book Techniques 155
Guidelines for Planning an Open-Book Questions and Exam 158
Summary 159
For practice 161

**CHAPTER 8  PERFORMANCE-BASED ASSESSMENT** 162

Performance Tests: Direct Measures of Competence 162
Performance Tests Can Assess Processes and Products 163
Performance Tests Can Be Embedded in Lessons 163
Performance Tests Can Assess Affective and Social Skills 165
Developing Performance Tests for Your Learners 166
Step 1: Deciding What to Test 166
Step 2: Designing the Assessment Context 170
Step 3: Specifying the Scoring Rubrics 172
Step 4: Specifying Testing Constraints 178
A Final Word 179
Summary 179
For Discussion and Practice 180
CHAPTER 9 PORTFOLIO ASSESSMENT 181
Rationale for the Portfolio 182
Ensuring Validity of the Portfolio 182
Developing Portfolio Assessments 183
  Step 1: Deciding on the Purposes for a Portfolio 183
  Step 2: Identifying Cognitive Skills and Dispositions 184
  Step 3: Deciding Who Will Plan the Portfolio 184
  Step 4: Deciding Which Products to Put in the Portfolio and How Many Samples of Each Product 184
  Step 5: Building the Portfolio Rubrics 185
  Step 6: Developing a Procedure to Aggregate All Portfolio Ratings 190
  Step 7: Determining the Logistics 193
Summary 196
For Practice 197

CHAPTER 10 ADMINISTERING, ANALYZING, AND IMPROVING THE TEST OR ASSESSMENT 198
Assembling the Test 198
  Packaging the Test 199
  Reproducing the Test 201
Administering the Test 201
Scoring the Test 203
Analyzing the Test 204
  Quantitative Item Analysis 204
  Qualitative Item Analysis 211
  Item Analysis Modifications for the Criterion-Referenced Test 212
Debriefing 216
  Debriefing Guidelines 217
The Process of Evaluating Classroom Achievement 218
Summary 220
For Practice 221

CHAPTER 11 MARKS AND MARKING SYSTEMS 222
What Is the Purpose of a Mark? 222
  Why Be Concerned about Marking? 222
  What Should a Mark Reflect? 223
Marking Systems 224
  Types of Comparisons 224
  Which System Should You Choose? 228
  Types of Symbols 229
Combining and Weighting the Components of a Mark 230
  Who Is the Better Teacher? 231

CONTENTS xi
Combining Grades from Quizzes, Tests, Papers, Homework, Etc., into a Single Mark 232
Practical Approaches to Equating Before Weighting in the Busy Classroom 236
Front-End Equating 236
Back-End Equating 238
Summary 240
For Practice 240

CHAPTER 12 SUMMARIZING DATA AND MEASURES OF CENTRAL TENDENCY 242
What Are Statistics? 242
Why Use Statistics? 243
Tabulating Frequency Data 244
  The List 244
  The Simple Frequency Distribution 245
  The Grouped Frequency Distribution 245
Steps in Constructing a Grouped Frequency Distribution 247
Graphing Data 250
  The Bar Graph, or Histogram 251
  The Frequency Polygon 251
  The Smooth Curve 253
Measures of Central Tendency 257
  The Mean 258
  The Median 259
  The Mode 264
  The Measures of Central Tendency in Various Distributions 266
Summary 267
For Practice 269

CHAPTER 13 VARIABILITY, THE NORMAL DISTRIBUTION, AND CONVERTED SCORES 270
The Range 270
The Semi-Interquartile Range (SIQR) 271
The Standard Deviation 272
  The Deviation Score Method for Computing the Standard Deviation 276
  The Raw Score Method for Computing the Standard Deviation 277
The Normal Distribution 279
  Properties of the Normal Distribution 280
Converted Scores 284
  z-Scores 286
  T-Scores 291
Summary 292
For Practice 292


<table>
<thead>
<tr>
<th>CHAPTER 14</th>
<th>CORRELATION</th>
<th>294</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Correlation Coefficient</td>
<td>295</td>
<td></td>
</tr>
<tr>
<td>Strength of a Correlation</td>
<td>296</td>
<td></td>
</tr>
<tr>
<td>Direction of a Correlation</td>
<td>296</td>
<td></td>
</tr>
<tr>
<td>Scatterplots</td>
<td>297</td>
<td></td>
</tr>
<tr>
<td>Where Does ( r ) Come From?</td>
<td>299</td>
<td></td>
</tr>
<tr>
<td>Causality</td>
<td>300</td>
<td></td>
</tr>
<tr>
<td>Other Interpretive Cautions</td>
<td>302</td>
<td></td>
</tr>
<tr>
<td>Summary</td>
<td>304</td>
<td></td>
</tr>
<tr>
<td>For Practice</td>
<td>305</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER 15</th>
<th>VALIDITY EVIDENCE</th>
<th>306</th>
</tr>
</thead>
<tbody>
<tr>
<td>Why Evaluate Tests?</td>
<td>306</td>
<td></td>
</tr>
<tr>
<td>Types of Validity Evidence</td>
<td>306</td>
<td></td>
</tr>
<tr>
<td>Content Validity Evidence</td>
<td>307</td>
<td></td>
</tr>
<tr>
<td>Criterion-Related Validity Evidence</td>
<td>307</td>
<td></td>
</tr>
<tr>
<td>Construct Validity Evidence</td>
<td>309</td>
<td></td>
</tr>
<tr>
<td>What Have We Been Saying? A Review</td>
<td>310</td>
<td></td>
</tr>
<tr>
<td>Interpreting Validity Coefficients</td>
<td>311</td>
<td></td>
</tr>
<tr>
<td>Content Validity Evidence</td>
<td>311</td>
<td></td>
</tr>
<tr>
<td>Concurrent and Predictive Validity Evidence</td>
<td>312</td>
<td></td>
</tr>
<tr>
<td>Summary</td>
<td>316</td>
<td></td>
</tr>
<tr>
<td>For Practice</td>
<td>317</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER 16</th>
<th>RELIABILITY</th>
<th>318</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methods of Estimating Reliability</td>
<td>318</td>
<td></td>
</tr>
<tr>
<td>Test–Retest or Stability</td>
<td>318</td>
<td></td>
</tr>
<tr>
<td>Alternate Forms or Equivalence</td>
<td>320</td>
<td></td>
</tr>
<tr>
<td>Internal Consistency</td>
<td>320</td>
<td></td>
</tr>
<tr>
<td>Interpreting Reliability Coefficients</td>
<td>323</td>
<td></td>
</tr>
<tr>
<td>Summary</td>
<td>326</td>
<td></td>
</tr>
<tr>
<td>For Practice</td>
<td>327</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER 17</th>
<th>ACCURACY AND ERROR</th>
<th>328</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error—What Is It?</td>
<td>328</td>
<td></td>
</tr>
<tr>
<td>The Standard Error of Measurement</td>
<td>330</td>
<td></td>
</tr>
<tr>
<td>Using the Standard Error of Measurement</td>
<td>331</td>
<td></td>
</tr>
<tr>
<td>More Applications</td>
<td>334</td>
<td></td>
</tr>
<tr>
<td>Standard Deviation or Standard Error of Measurement?</td>
<td>336</td>
<td></td>
</tr>
<tr>
<td>Why All the Fuss About Error?</td>
<td>337</td>
<td></td>
</tr>
<tr>
<td>Error within Test Takers</td>
<td>337</td>
<td></td>
</tr>
<tr>
<td>Error within the Test</td>
<td>337</td>
<td></td>
</tr>
<tr>
<td>Error in Test Administration</td>
<td>338</td>
<td></td>
</tr>
<tr>
<td>Error in Scoring</td>
<td>338</td>
<td></td>
</tr>
<tr>
<td>Summary</td>
<td>390</td>
<td></td>
</tr>
<tr>
<td>For Practice</td>
<td>392</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER 18</th>
<th>STANDARDIZED TESTS</th>
<th>349</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sources of Error Influencing Various Reliability Coefficients</td>
<td>339</td>
<td></td>
</tr>
<tr>
<td>Test–Retest</td>
<td>339</td>
<td></td>
</tr>
<tr>
<td>Alternate Forms</td>
<td>339</td>
<td></td>
</tr>
<tr>
<td>Internal Consistency</td>
<td>340</td>
<td></td>
</tr>
<tr>
<td>Band Interpretation</td>
<td>341</td>
<td></td>
</tr>
<tr>
<td>Steps: Band Interpretation</td>
<td>342</td>
<td></td>
</tr>
<tr>
<td>A Final Word</td>
<td>346</td>
<td></td>
</tr>
<tr>
<td>Summary</td>
<td>346</td>
<td></td>
</tr>
<tr>
<td>For Practice</td>
<td>347</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER 19</th>
<th>STANDARDIZED TESTS</th>
<th>349</th>
</tr>
</thead>
<tbody>
<tr>
<td>What Is a Standardized Test?</td>
<td>349</td>
<td></td>
</tr>
<tr>
<td>Do Test Stimuli, Administration, and Scoring Have to Be Standardized?</td>
<td>351</td>
<td></td>
</tr>
<tr>
<td>Standardized Testing: Effects of Accommodations and Alternate Assessments</td>
<td>351</td>
<td></td>
</tr>
<tr>
<td>Uses of Standardized Achievement Tests</td>
<td>352</td>
<td></td>
</tr>
<tr>
<td>Will Performance and Portfolio Assessment Make Standardized Tests Obsolete?</td>
<td>353</td>
<td></td>
</tr>
<tr>
<td>Administering Standardized Tests</td>
<td>354</td>
<td></td>
</tr>
<tr>
<td>Types of Scores Offered for Standardized Achievement Tests</td>
<td>356</td>
<td></td>
</tr>
<tr>
<td>Grade Equivalents</td>
<td>356</td>
<td></td>
</tr>
<tr>
<td>Age Equivalents</td>
<td>357</td>
<td></td>
</tr>
<tr>
<td>Percentile Ranks</td>
<td>358</td>
<td></td>
</tr>
<tr>
<td>Standard Scores</td>
<td>359</td>
<td></td>
</tr>
<tr>
<td>Interpreting Standardized Tests: Test and Student Factors</td>
<td>360</td>
<td></td>
</tr>
<tr>
<td>Test-Related Factors</td>
<td>360</td>
<td></td>
</tr>
<tr>
<td>Student-Related Factors</td>
<td>367</td>
<td></td>
</tr>
<tr>
<td>Aptitude–Achievement Discrepancies</td>
<td>371</td>
<td></td>
</tr>
<tr>
<td>Interpreting Standardized Tests: Parent-Teacher Conferences and Educational Decision Making</td>
<td>375</td>
<td></td>
</tr>
<tr>
<td>An Example: Pressure to Change an Educational Placement</td>
<td>375</td>
<td></td>
</tr>
<tr>
<td>A Second Example: Pressure from the Opposite Direction</td>
<td>380</td>
<td></td>
</tr>
<tr>
<td>Interpreting Standardized Tests: Score Reports from Publishers</td>
<td>383</td>
<td></td>
</tr>
<tr>
<td>The Press-On Label</td>
<td>386</td>
<td></td>
</tr>
<tr>
<td>A Criterion-Referenced Skills Analysis or Mastery Report</td>
<td>387</td>
<td></td>
</tr>
<tr>
<td>An Individual Performance Profile</td>
<td>388</td>
<td></td>
</tr>
<tr>
<td>Other Publisher Reports and Services</td>
<td>389</td>
<td></td>
</tr>
<tr>
<td>Summary</td>
<td>390</td>
<td></td>
</tr>
<tr>
<td>For Practice</td>
<td>392</td>
<td></td>
</tr>
</tbody>
</table>
CHAPTER 19  TYPES OF STANDARDIZED TESTS  393

Standardized Achievement Tests  393
  Achievement Test Batteries, or Survey Batteries  394
  Single-Subject Achievement Tests  395
  Diagnostic Achievement Tests  396

Standardized Academic Aptitude Tests  397
  The History of Academic Aptitude Testing  397

Stability of IQ Scores  398
  What Do IQ Tests Predict?  398
  Individually Administered Academic Aptitude Tests  400
  Group Administered Academic Aptitude Tests  401

Standardized Personality Assessment Instruments  401
  What Is Personality?  402
  Objective Personality Tests  403
  Projective Personality Tests  404

Summary  404
  For Discussion  405

CHAPTER 20  ASSESSING CHILDREN WITH DISABILITIES IN THE REGULAR CLASSROOM: A GROWING NEED  393

A Brief History of Special Education  410
  P.L. 94–142 and the Individuals with Disabilities Education Act (IDEA)  410
  Section 504 of the Rehabilitation Act  411

Special Education Service Delivery: An Evolution  412
  Service Delivery Setting  412
  Determining Eligibility for Services  414
  Learning Disability Determination: Discrepancy or RTI?  415

Disability Categories to Developmental Delays  418

IDEA–04 And The Classroom Teacher  419
  Testing or the Assessment Process?  420
  Child Identification  422
  Individual Assessment  423
  Individual Educational Plan (IEP) Development  425
  Individualized Instruction  428
  Reviewing the IEP  429

At the Other End of the Curve: The Gifted and Talented Child  430
  Defining “Gifted and Talented”  430
  Assessment and Identification  430

Summary  435
  For Discussion  437

CHAPTER 21  METHODS TO ASSESS CHILDREN WITH DISABILITIES IN REGULAR EDUCATION CLASSROOMS  438

IDEA and NCLB: Issues and Questions  439
  Assistance for Teachers  439
  Can Regular Teachers Assess Children With Disabilities?  440
  Should Regular Teachers Assess Children With Disabilities?  440
  Assessing Academic Performance and Progress  441
  Teacher-Made Tests and Assessments  441
  Standardized Tests and Assessments  442
  Limitations of Accommodations and Alternate Assessments  442

Assessing Behavioral and Attitudinal Factors  443
  Assessment, Not Diagnosis  443
  Classroom Diversity, Behavior, and Attitudes  444

Monitoring Children with Disabilities Who Are Taking Medication  459
  Medication Use Is Increasing  460
  Side Effects Are Always Present  460
  The Teacher’s Role in Evaluating Medication and Psychosocial Interventions  460

Commonly Used Standardized Scales and Checklists  461

Summary  465
  For Discussion  467

CHAPTER 22  IN THE CLASSROOM: A SUMMARY DIALOGUE  469

High-Stakes Testing and NCLB  474
  Criterion-Referenced versus Norm-Referenced Tests  474
  New Responsibilities for Teachers Under IDEA  475

Instructional Objectives  476
  The Test Blueprint  476
  Essay Items and the Essay Scoring Guides  477
  Reliability, Validity Evidence, and Test Statistics  477

Grades and Marks  479
  Some Final Thoughts  480