Contents

PART I. INSTRUMENT DEVELOPMENT AND ANALYSIS

1 • Introduction 3

Problems in Social Science Measurement 4
What Is Measurement Theory? 5
Measurement Defined 6
The Nominal Level of Measurement 6
The Ordinal Level of Measurement 7
The Interval Level of Measurement 7
The Ratio Level of Measurement 8
Criticisms of Stevens's Levels of Measurement 8
A Brief History of Testing 9
The Chinese Civil Service Examinations 9
Testing in Ancient Greece 11
Early European Testing 12
Testing in the United States 15
Testing in Business and Industry 18
Personality Assessment 18
Summary 19
Exercises 20

2 • Norms and Standardized Scores 21

Which to Use? 22
Norm Groups 23
Important Characteristics of the Norm Group: The "Three R's" 23

Types of Norm-Referenced Scores 24
Percentile Ranks 25
Standardized and Normalized Scores 28
Stanines 32
Normal Curve Equivalents 33
Developmental-Level Scores 33
Criterion-Referenced Testing 37
Summary 37
Exercises 38

3 • The Test Development Process 41

Steps in Scale Development 42
State the Purpose of the Scale 43
Identify and Define the Domain 43
Determine Whether a Measure Already Exists 44
Determine the Item Format 46
Write Out the Testing Objectives 46
Create the Initial Item Pool 50
Conduct the Initial Item Review 52
Conduct Preliminary Item Tryouts 53
Conduct a Large-Scale Field Test of Items 54
Prepare Guidelines for Administration 57
Summary 59
Exercises 60

4 • Writing Cognitive Items 63

Objective Item Types 64
Multiple-Choice Items 64
True–False Items 66
Matching Items 67
Short-Answer or Completion Items 69
Performance Assessments 71
Essay Questions 72
Performance Tasks 76
Summary 81
Exercises 82

5 • Writing Noncognitive Items 85

Noncognitive Item Types 86
Thurstone Scaling 86
Likert Scaling 89
Guttman Scaling 94
Theories of Item Responding 95
The Cognitive Process Model of Responding 95
Item Responses as Social Encounters 102

Problems in Measuring Noncognitive Outcomes 104
Response Distortion 104
Managing Response Distortion 108
Practical Issues in Noncognitive Scale Construction 111
Number of Scale Points 111
Labeling of Response Options 112
Inclusion of Negatively Oriented Items 113
Including a Neutral Option 115
Summary 116
Exercises 117

6 • Item Analysis for Cognitive and Noncognitive Items 120

Item Analysis for Cognitive Items 121
Item Difficulty 122
Item Discrimination 123
Evaluating the Distractors for Multiple-Choice Items 126
Corrections for Guessing 128
Summary of Analyses for Cognitive Items 130
Item Analysis for Noncognitive Items 131
Frequency Distributions and Descriptive Statistics 131
Interitem Correlations 135
Item–Total Correlations and Information from Reliability Analyses 137
Group Comparisons 139
Factor Analytic Methods 141
Summary of Analyses for Noncognitive Items 143
Use of Item Analysis Information 144
Exercises 146

PART II. RELIABILITY AND VALIDITY

7 • Introduction to Reliability and the Classical Test Theory Model 155

What Is Reliability? 155
Measurement Error and CTT 158
More on CTT 160
Properties of True and Error Scores in CTT 161
The CTT Definition of Reliability 162
Correlation between True and Observed Scores:
The Reliability Index 163
Parallel, Tau-Equivalent, and Congeneric Measures 164
Reliability as the Correlation between Scores
on Parallel Tests 167
Summary 169
Exercises 170

8 • Methods of Assessing Reliability 172

Internal Consistency 172
Reliability of a Composite 173
The Spearman–Brown Prophecy Formula and Split-Half Reliability 175
Coefficient Alpha 176
Other Internal Consistency Coefficients 183
Recommended Values for Internal Consistency Indices 184
Factors Affecting Internal Consistency Coefficient Values 185
Computational Examples for Coefficient Alpha 188
Test–Retest Reliability 190
Factors Affecting Coefficients of Stability 192
Recommended Values for Coefficients of Stability 194
Alternate Forms Reliability 195
Factors Affecting Coefficients of Equivalence 196
Recommended Values for Coefficients of Equivalence 196
Combining Alternate Forms and Test–Retest Reliability 197
Factors Affecting Coefficients of Equivalence and Stability 197
Recommended Values for Coefficients of Equivalence and Stability 197
The Standard Error of Measurement 198
Factors Affecting the SEM 199
Using the SEM to Place Confidence Intervals around Scores 199
Sample Dependence of Reliability Coefficients and the SEM 200
Reliability of Difference Scores 201
Summary 205
Exercises 206

9 • Interrater Agreement and Reliability 210

Measures of Interrater Agreement 212
Nominal Agreement 212
Cohen's Kappa 213
Measures of Interrater Reliability 215
Coefficient Alpha 215
Intraclass Correlation 215
Summary 221
Exercises 224

10 • Generalizability Theory 226

Basic Concepts and Terminology 226
Facets, Objects of Measurement, and Universe Scores 227
Crossed and Nested Facets 229
Random and Fixed Facets 229
G Studies and D Studies 230
The G Theory Model 230
Computation of Variance Components 232
Computation of Variance Components for a One-Facet Design 233

Computation of Variance Components for a Two-Facet Design 236
Variance Components for Nested Designs 238
Variance Components for Designs with Fixed Facets 240
Decision Studies 244
Relative and Absolute Interpretations 245
Calculating the G and Phi Coefficients 245
Use of the D Study to Determine the Optimal Test Design 247
Decision Studies with Nested or Fixed Facets 249
Summary 250
Exercises 252

11 • Validity 254

Validity Defined 255
Traditional Forms of Validity Evidence:
A Historical Perspective 255
Original Validity Types 256
Arguments against the "Tripartite" View of Validity 259
Current Conceptualizations of Validity 260
The Unified View of Validity 260
Focus on Interpretation and Use of Test Scores 261
Focus on Explanation and Cognitive Models 263
Inclusion of Values and Test Consequences in the Validity Framework 264
Obtaining Evidence of Validity 265
Introduction to the Argument-Based Approach to Validity 266
Types of Validity Evidence 267
Summary 295
Exercises 296

PART III. ADVANCED TOPICS IN MEASUREMENT THEORY

12 • Exploratory Factor Analysis 301

The EFA–CFA Distinction 302
The EFA Model 303
The EFA Model: Diagrammatic Form 304
The EFA Model: Equation Form 307
Steps in Conducting EFA 310
Extracting the Factors 311
Determining the Number of Factors to Retain 318
Rotating the Factors 327
Interpreting the Factors 338
Data Requirements for EFA 341
Sample-Size Requirements 345
Summary 345
Exercises 346

13 • Confirmatory Factor Analysis 350

Differences between Exploratory and Confirmatory Factor Analyses 350
Advantages of CFA 351
CFA Model and Equations 352
Steps in Conducting a CFA 357
Model Specification 359
Model Identification 362
Estimation of Model Parameters 366
Model Testing 376
Respecification of the Model 382
Data Preparation and CFA Assumptions 388
Normality of Variable Distributions 389
Variable Scales 389
Outliers 389
Missing Data 390
Sampling Method 391
Sample Size 391
CFA-Based Reliability Estimation 392
Tests of Parameter Estimate Equivalence 392
Calculation of Coefficient Omega 395
Summary 397
Exercises 398

14 • Item Response Theory 403
WITH CHRISTINE E. DEMARS

Item Response Functions for IRT 405
IRT Models 406
The One-Parameter Logistic Model 407
The Two-Parameter Logistic Model 410
The Three-Parameter Logistic Model 410
IRT Models for Polytomous Items 413
Indeterminacy and Scaling 416
Scaling for the Rasch Model 417
Scaling for the 2PL and 3PL Models 417
Invariance of Parameter Estimates 418
Estimation 419
Maximum Likelihood Estimation 420
Bayesian Estimation Methods 423
Sample Size Requirements 428
Information, Standard Error of Measurement, and Reliability 429
Maximum Likelihood Estimation 429
EAP Estimation 432
IRT Assumptions 433
Correct Dimensionality 433
Local Independence 434
Functional Form 435

IRT Applications 437
Test Form Assembly 437
Equating 438
Computer Adaptive Testing Applications 439
Differential Item Functioning Applications 440
Summary 440
Exercises 441

15 • Diagnostic Classification Models 446
WITH LAINE P. BRADSHAW

Categorical Latent Variables for DCMs 447
When to Use DCMs 448
Attribute Profiles 448
Diagnostic Classification Model: A Confirmatory Latent Class Model 449
The Latent Class Model 450
IRFs for DCMs 451
The Log-Linear Cognitive Diagnosis Model: A General DCM 452
Link Functions for DCMs 452
The Q-Matrix 455
IRF for Complex Structure Items 461
Fully Extending the IRF for the LCDM 465
Other General DCMs 467
Submodels of the LCDM 468
The Deterministic Inputs Noisy And Gate Model 468
The Compensatory Reparameterized Unified Model 468
The Deterministic Inputs Noisy Or Gate Model 469
Other Models 469
Which Model Should I Use? 470
Examinee Classifications 471
Summary 475
Exercises 476

16 • Bias, Fairness, and Legal Issues in Testing 478

Impact, Item and Test Bias, Differential Item Functioning, and *Fairness* Defined 479
Detecting Test and Item Bias 480
Test Bias 480
Item Bias 483
Choosing a DIF Detection Method 498
Purification of the Matching Variable 499
Interpretation of DIF and Test Bias 499
DIF as Construct-Irrelevant Variance 500
Sources of Test Bias 500
Test Fairness 502
Universal Design 503
Accommodations and Modifications 503

Need for More Research on DIF and Test Bias 505
Sensitivity Reviews 505
Legal Issues in Testing 506
Legislation under Which Tests Can Be Challenged 507
Court Cases Relevant to Testing 508
Summary 514
Exercises 516

17 • Standard Setting 519

Common Elements of Standard-Setting Procedures 522
Step 1: Select a Standard-Setting Procedure 522
Step 2: Choose the Panelists 523
Step 3: Prepare Descriptions of Each Performance Category 524
Step 4: Train the Panelists to Use the Chosen Procedure 524
Step 5: Collect Panelists' Judgments 525
Step 6: Provide Panelists with Feedback and Discuss 525
Step 7: Collect a Second Set of Judgments and Create Recommended Cut Scores 527
Step 8: Conduct an Evaluation of the Standard-Setting Process 527
Step 9: Compile a Technical Report, Including Validity Evidence 528
Standard-Setting Procedures 528
The Angoff Method 529
The Bookmark Procedure 531
The Contrasting Groups Method 533
The Borderline Group Method 535
The Body of Work Method 536
Validity Evidence for Standard Setting 539
Procedural Evidence 540
Internal Evidence 541
External Evidence 542
Summary 544
Exercises 545

18 • Test Equating 547

Equating Defined 548
Alternatives to Equating 550
Equating Designs 551
Single-Group Design 551
Random-Groups Design 552
Common Item Nonequivalent Groups Design 552
Methods of Equating 554
Mean Equating 554
Linear Equating 555
Equipercentile Equating 560
IRT Equating Methods 570

Practical Considerations in Equating 576
Guidelines for Choosing Common Items 576
Error in Equating 577
Sample-Size Requirements 578
Systematic Equating Error 579
Choice of Equating Method 580
Summary 581
Exercises 582

Answers to Exercises 585
References 623
Author Index 645
Subject Index 651
About the Author 661