Contents

Preface **vii**

Notation **xxi**

1 Introduction and Concepts **1**
1.1 Equating and Related Concepts 1
1.1.1 Test Forms and Test Specifications 2
1.1.2 Equating 2
1.1.3 Processes That Are Related to Equating 3
1.1.4 Equating and Score Scales 4
1.1.5 Equating and the Test Score Decline of the 1960s and 1970s 7
1.2 Equating and Scaling in Practice—A Brief Overview of This Book 7
1.3 Properties of Equating 9
1.3.1 Symmetry Property 10
1.3.2 Same Specifications Property 10
1.3.3 Equity Properties 10
1.3.4 Observed Score Equating Properties 12
1.3.5 Group Invariance Property 13
1.4 Equating Designs 13
1.4.1 Random Groups Design 13
1.4.2 Single Group Design 15
1.4.3 Single Group Design with Counterbalancing 15

1.4.4 ASVAB Problems with a Single Group Design . . . 17
1.4.5 Common-Item Nonequivalent Groups Design 19
1.4.6 NAEP Reading Anomaly—Problems with Common Items 22
1.5 Error in Estimating Equating Relationships 23
1.6 Evaluating the Results of Equating 24
1.7 Testing Situations Considered 25
1.8 Preview 26
1.9 Exercises 27

2 Observed Score Equating Using the Random Groups Design 29
2.1 Mean Equating 30
2.2 Linear Equating 31
2.3 Properties of Mean and Linear Equating 32
2.4 Comparison of Mean and Linear Equating 34
2.5 Equipercentile Equating 36
2.5.1 Graphical Procedures 39
2.5.2 Analytic Procedures 43
2.5.3 Properties of Equated Scores in Equipercentile Equating 46
2.6 Estimating Observed Score Equating Relationships 48
2.7 Scale Scores 52
2.7.1 Linear Conversions 54
2.7.2 Truncation of Linear Conversions 55
2.7.3 Nonlinear Conversions 56
2.8 Equating Using Single Group Designs 62
2.9 Equating Using Alternate Scoring Schemes 63
2.10 Preview of What Follows 64
2.11 Exercises 64

3 Random Groups—Smoothing in Equipercentile Equating 67
3.1 A Conceptual Statistical Framework for Smoothing 68
3.2 Properties of Smoothing Methods 72
3.3 Presmoothing Methods 73
3.3.1 Polynomial Log-Linear Method 74
3.3.2 Strong True Score Method 75
3.3.3 Illustrative Example 77
3.4 Postsmoothing Methods 84
3.4.1 Illustrative Example 89
3.5 Practical Issues in Equipercentile Equating 91
3.5.1 Summary of Smoothing Strategies 91
3.5.2 Equating Error and Sample Size 98
3.6 Exercises 100

4 Nonequivalent Groups—Linear Methods 103
4.1 Tucker Method . . . 105
4.1.1 Linear Regression Assumptions . . . 106
4.1.2 Conditional Variance Assumptions . . . 106
4.1.3 Intermediate Results . . . 107
4.1.4 Final Results . . . 108
4.1.5 Special Cases . . . 108
4.2 Levine Observed Score Method . . . 109
4.2.1 Correlational Assumptions . . . 109
4.2.2 Linear Regression Assumptions . . . 110
4.2.3 Error Variance Assumptions . . . 110
4.2.4 Intermediate Results . . . 111
4.2.5 General Results . . . 111
4.2.6 Classical Congeneric Model Results . . . 112
4.3 Levine True Score Method . . . 115
4.3.1 Results . . . 116
4.3.2 First-Order Equity . . . 118
4.4 Illustrative Example and Other Topics . . . 120
4.4.1 Illustrative Example . . . 121
4.4.2 Synthetic Population Weights . . . 124
4.4.3 Mean Equating . . . 125
4.4.4 Decomposing Observed Differences in Means and Variances . . . 125
4.4.5 Relationships Among Tucker and Levine Equating Methods . . . 128
4.4.6 Scale Scores . . . 130
4.5 Appendix: Proof $\sigma_s^2(T_X) = \gamma_1^2\sigma_s^2(T_V)$ Under Classical Congeneric Model . . . 131
4.6 Exercises . . . 132

5 Nonequivalent Groups—Equipercentile Methods 135
5.1 Frequency Estimation Equipercentile Equating . . . 135
5.1.1 Conditional Distributions . . . 136
5.1.2 Frequency Estimation Method . . . 136
5.1.3 Evaluating the Frequency Estimation Assumption . . . 138
5.1.4 Numerical Example . . . 139
5.1.5 Estimating the Distributions . . . 142
5.2 Braun-Holland Linear Method . . . 144
5.3 Chained Equipercentile Equating . . . 145
5.4 Illustrative Example . . . 147
5.4.1 Illustrative Results . . . 147
5.4.2 Comparison Among Methods . . . 151
5.4.3 Practical Issues in Equipercentile Equating with Common Items . . . 152
5.5 Exercises . . . 153

6 Item Response Theory Methods **155**
6.1 Some Necessary IRT Concepts . . . 156
6.1.1 Unidimensionality and Local Independence Assumptions . . . 156
6.1.2 IRT Models . . . 157
6.1.3 IRT Parameter Estimation . . . 160
6.2 Transformations of IRT Scales . . . 161
6.2.1 Transformation Equations . . . 162
6.2.2 Demonstrating the Appropriateness of Scale Transformations . . . 162
6.2.3 Expressing A and B Constants . . . 163
6.2.4 Expressing A and B Constants in Terms of Groups of Items and/or Persons . . . 164
6.3 Transforming IRT Scales When Parameters Are Estimated 165
6.3.1 Designs . . . 166
6.3.2 Mean/Sigma and Mean/Mean Transformation Methods . . . 167
6.3.3 Characteristic Curve Transformation Methods . . . 168
6.3.4 Comparisons Among Scale Transformation Methods 173
6.4 Equating and Scaling . . . 175
6.5 Equating True Scores . . . 176
6.5.1 Test Characteristic Curves . . . 176
6.5.2 True Score Equating Process . . . 176
6.5.3 The Newton-Raphson Method . . . 177
6.5.4 Using True Score Equating with Observed Scores . . 180
6.6 Equating Observed Scores . . . 181
6.7 IRT True Score Versus IRT Observed Score Equating . . . 184
6.8 Illustrative Example . . . 185
6.8.1 Item Parameter Estimation and Scaling . . . 185
6.8.2 IRT True Score Equating . . . 191
6.8.3 IRT Observed Score Equating . . . 194
6.8.4 Rasch Equating . . . 198
6.9 Using IRT Calibrated Item Pools . . . 201
6.9.1 Common-Item Equating to a Calibrated Pool . . . 201
6.9.2 Item Preequating . . . 205
6.9.3 Robustness to Violations of IRT Assumptions . . . 207
6.10 Equating with Polytomous IRT . . . 208
6.10.1 Polytomous IRT Models for Ordered Responses . . . 209
6.10.2 Scoring Function, Item Response Function, and Test Characteristic Curve . . . 214
6.10.3 Parameter Estimation and Scale Transformation with Polytomous IRT Models . . . 215
6.10.4 True Score Equating . . . 219
6.10.5 Observed Score Equating . . . 219
6.10.6 Example Using the Graded Response Model . . . 220

6.11 Practical Issues and Caveat 227
6.12 Exercises 228

7 Standard Errors of Equating 231
7.1 Definition of Standard Error of Equating 232
7.2 The Bootstrap 235
7.2.1 Standard Errors Using the Bootstrap 235
7.2.2 Standard Errors of Equating 236
7.2.3 Parametric Bootstrap 238
7.2.4 Standard Errors of Equipercentile Equating with Smoothing 240
7.2.5 Standard Errors of Scale Scores 241
7.2.6 Standard Errors of Equating Chains 242
7.2.7 Mean Standard Error of Equating 243
7.2.8 Caveat 244
7.3 The Delta Method 245
7.3.1 Mean Equating Using Single Group and Random Groups Designs 246
7.3.2 Linear Equating Using the Random Groups Design . 247
7.3.3 Equipercentile Equating Using the Random Groups Design 248
7.3.4 Standard Errors for Other Designs 249
7.3.5 Approximations 251
7.3.6 Standard Errors for Scale Scores 253
7.3.7 Standard Errors of Equating Chains 254
7.3.8 Using Delta Method Standard Errors 255
7.4 Using Standard Errors in Practice 261
7.5 Exercises 263

8 Practical Issues in Equating 267
8.1 Equating and the Test Development Process 269
8.1.1 Test Specifications 269
8.1.2 Characteristics of Common-Item Sets 271
8.1.3 Changes in Test Specifications 272
8.2 Data Collection: Design and Implementation 273
8.2.1 Choosing Among Equating Designs 273
8.2.2 Developing Equating Linkage Plans 277
8.2.3 Examinee Groups Used in Equating 285
8.2.4 Sample Size Requirements 288
8.3 Choosing from Among the Statistical Procedures 290
8.3.1 Equating Criteria in Research Studies 290
8.3.2 Characteristics of Equating Situations 292
8.4 Choosing from Among Equating Results 296
8.4.1 Equating Versus Not Equating 296
8.4.2 Use of Robustness Checks 296

8.4.3 Choosing from Among Results in the Random Groups Design . . . 297
8.4.4 Choosing from Among Results in the Common-Item Nonequivalent Groups Design . . . 298
8.4.5 Use of Consistency Checks . . . 298
8.4.6 Equating and Score Scales . . . 300
8.4.7 Assessing First- and Second-Order Equity for Scale Scores . . . 301
8.5 Importance of Standardization Conditions and Quality Control Procedures . . . 306
8.5.1 Test Development . . . 307
8.5.2 Test Administration and Standardization Conditions 307
8.5.3 Quality Control . . . 309
8.5.4 Reequating . . . 310
8.6 Conditions Conducive to Satisfactory Equating . . . 312
8.7 Comparability Issues in Special Circumstances . . . 312
8.7.1 Comparability Issues with Computer-Based Tests . . 314
8.7.2 Comparability of Performance Assessments . . . 320
8.7.3 Score Comparability with Optional Test Sections . . 323
8.8 Conclusion . . . 324
8.9 Exercises . . . 325

9 Score Scales 329
9.1 Scaling Perspectives . . . 331
9.2 Score Transformations . . . 336
9.3 Incorporating Normative Information . . . 337
9.3.1 Linear Transformations . . . 337
9.3.2 Nonlinear Transformations . . . 338
9.3.3 Example: Normalized Scale Scores . . . 340
9.3.4 Importance of Norm Group in Setting the Score Scale 344
9.4 Incorporating Score Precision Information . . . 345
9.4.1 Rules of Thumb for Number of Distinct Score Points 345
9.4.2 Linearly Transformed Score Scales with a Given Standard Error of Measurement . . . 348
9.4.3 Score Scales with Approximately Equal Conditional Standard Errors of Measurement . . . 348
9.4.4 Example: Incorporating Score Precision . . . 351
9.4.5 Evaluating Psychometric Properties of Scale Scores . 354
9.4.6 The IRT θ-Scale as a Score Scale . . . 358
9.5 Incorporating Content Information . . . 358
9.5.1 Item Mapping . . . 358
9.5.2 Scale Anchoring . . . 361
9.5.3 Standard Setting . . . 361
9.5.4 Numerical Example . . . 364
9.5.5 Practical Usefulness . . . 366

9.6 Maintaining Score Scales 366
9.7 Scales for Test Batteries and Composites 368
9.7.1 Test Batteries 368
9.7.2 Composite Scores 369
9.7.3 Maintaining Scales for Batteries and Composites . . 371
9.8 Vertical Scaling and Developmental Score Scales 372
9.8.1 Structure of Batteries 373
9.8.2 Type of Domain Being Measured 375
9.8.3 Definition of Growth 376
9.8.4 Designs for Data Collection for Vertical Scaling . . . 377
9.8.5 Test Scoring 381
9.8.6 Hieronymus Statistical Methods 381
9.8.7 Thurstone Statistical Methods 383
9.8.8 IRT Statistical Methods 387
9.8.9 Thurstone Illustrative Example 393
9.8.10 IRT Illustrative Example 401
9.8.11 Statistics for Comparing Scaling Results 410
9.8.12 Some Limitations of Vertically Scaled Tests 412
9.8.13 Research on Vertical Scaling 414
9.9 Exercises 418

10 Linking **423**
10.1 Linking Categorization Schemes and Criteria 424
10.1.1 Types of Linking 427
10.1.2 Mislevy/Linn Taxonomy 429
10.1.3 Degrees of Similarity 433
10.2 Group Invariance 437
10.2.1 Statistical Methods Using Observed Scores 437
10.2.2 Statistics for Overall Group Invariance 441
10.2.3 Statistics for Pairwise Group Invariance 443
10.2.4 Example: ACT and ITED Science Tests 444
10.3 Additional Examples 465
10.3.1 Extended Time 465
10.3.2 Test Adaptations and Translated Tests 467
10.4 Discussion 469
10.5 Exercises 470

11 Current and Future Challenges **473**
11.1 Score Scales 473
11.2 Equating 474
11.3 Vertical Scaling 475
11.4 Linking 475
11.5 Summary 476

References **477**

Appendix A: Answers to Exercises **511**

Appendix B: Computer Programs **533**

Index **535**