

Contents

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	6
1.2	Equating and Scaling in Practice: A Brief Overview of This Book	7
1.3	Properties of Equating	8
1.3.1	Symmetry Property	9
1.3.2	Same Specifications Property	9
1.3.3	Equity Properties	9
1.3.4	Observed Score Equating Properties	11
1.3.5	Group Invariance Property	12
1.4	Equating Designs	12
1.4.1	Random Groups Design	13
1.4.2	Single Group Design	14
1.4.3	Single Group Design with Counterbalancing	14
1.4.4	ASVAB Problems with a Single Group Design	16
1.4.5	Common-Item Nonequivalent Groups Design	18
1.4.6	NAEP Reading Anomaly: Problems with Common Items	20
1.5	Error in Estimating Equating Relationships	21
1.6	Evaluating the Results of Equating	22
1.7	Testing Situations Considered	23
1.8	Preview	24
1.9	Exercises	25
	References	26
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.	33
2.5	Equipercentile Equating.	36
2.5.1	Graphical Procedures	38
2.5.2	Analytic Procedures	42
2.5.3	Properties of Equated Scores in Equipercentile Equating.	45
2.6	Estimating Observed Score Equating Relationships.	46
2.7	Scale Scores	50
2.7.1	Linear Conversions	50
2.7.2	Truncation of Linear Conversions.	53
2.7.3	Nonlinear Conversions	54
2.8	Equating Using Single Group Designs	60
2.9	Equating Using Alternate Scoring Schemes	60
2.10	Preview of What Follows	61
2.11	Exercises	62
	References	63
3	Random Groups: Smoothing in Equipercentile Equating	65
3.1	A Conceptual Statistical Framework for Smoothing	66
3.2	Properties of Smoothing Methods.	69
3.3	Presmoothing Methods	70
3.3.1	Polynomial Log-Linear Method	70
3.3.2	Strong True Score Method.	72
3.3.3	Illustrative Example	74
3.4	Postsmoothing Methods.	80
3.4.1	Illustrative Example	85
3.5	The Kernel Method of Equating.	89
3.6	Practical Issues in Equipercentile Equating	93
3.6.1	Summary of Smoothing Strategies	94
3.6.2	Smoothing and Population Distribution Irregularities.	95
3.6.3	Equating Error, Sample Size, and Smoothing Method	96
3.7	Exercises	98
	References	99
4	Nonequivalent Groups: Linear Methods	103
4.1	Tucker Method.	105
4.1.1	Linear Regression Assumptions	105
4.1.2	Conditional Variance Assumptions	106
4.1.3	Intermediate Results	107
4.1.4	Final Results	108
4.1.5	Special Cases	109

4.2	Levine Observed Score Method	109
4.2.1	Correlational Assumptions	110
4.2.2	Linear Regression Assumptions	110
4.2.3	Error Variance Assumptions.	111
4.2.4	Intermediate Results	111
4.2.5	General Results	112
4.2.6	Classical Congeneric Model Results	113
4.3	Levine True Score Method	116
4.3.1	Results	117
4.3.2	First-Order Equity.	119
4.4	Chained Linear Equating	121
4.4.1	Chained Linear Observed Score Equating	122
4.4.2	Chained Linear True Score Equating.	123
4.5	Illustrative Example and Other Topics	124
4.5.1	Illustrative Example	125
4.5.2	Synthetic Population Weights.	128
4.5.3	Mean Equating.	128
4.5.4	Decomposing Observed Differences in Means and Variances.	129
4.5.5	Relationships Among Linear Observed Score Methods	132
4.5.6	Relationships Involving Levine Methods	135
4.5.7	Other Issues Involving Methods	137
4.5.8	Scale Scores.	137
4.6	Appendix: Proof that $\sigma_s^2(T_X) = \gamma_1^2 \sigma_s^2(T_Y)$ Under the Classical Congeneric Model	139
4.7	Exercises	139
	References	141
5	Nonequivalent Groups: Equipercentile Methods	143
5.1	Frequency Estimation Method	143
5.1.1	Conditional Distributions	144
5.1.2	Assumptions and Procedures	144
5.1.3	Numerical Example.	147
5.1.4	Estimating the Distributions	150
5.1.5	Special Case: Braun-Holland Linear Method	151
5.1.6	Illustrative Example	152
5.2	Other Methods	158
5.2.1	Modified Frequency Estimation	158
5.2.2	Chained Equipercentile Equating	159
5.2.3	Illustrative Example	164
5.3	Practical Issues.	165
5.4	Exercises	166
	References	166

6	Item Response Theory Methods	171
6.1	Some Necessary IRT Concepts.	172
6.1.1	Unidimensionality and Local Independence Assumptions.	172
6.1.2	IRT Models	173
6.1.3	IRT Parameter Estimation	176
6.2	Transformations of IRT Scales	177
6.2.1	Transformation Equations	177
6.2.2	Demonstrating the Appropriateness of Scale Transformations	178
6.2.3	Expressing A and B Constants	179
6.2.4	Expressing A and B Constants in Terms of Groups of Items and/or Persons	180
6.3	Transforming IRT Scales When Parameters are Estimated.	181
6.3.1	Designs	182
6.3.2	Mean/Sigma and Mean/Mean Transformation Methods	183
6.3.3	Characteristic Curve Transformation Methods	184
6.3.4	Comparisons Among Scale Transformation Methods	189
6.4	Equating and Scoring	191
6.5	Equating True Scores	192
6.5.1	Test Characteristic Curves	192
6.5.2	True Score Equating Process	193
6.5.3	The Newton-Raphson Method	193
6.5.4	Using True Score Equating with Observed Scores	196
6.6	Equating Observed Scores	197
6.7	IRT True Score Versus IRT Observed Score Equating	201
6.8	Illustrative Example	201
6.8.1	Item Parameter Estimation and Scaling	202
6.8.2	IRT True Score Equating.	206
6.8.3	IRT Observed Score Equating	207
6.8.4	Rasch Equating	213
6.9	Using IRT Calibrated Item Pools and Other Designs	214
6.9.1	Common-Item Equating to a Calibrated Pool	215
6.9.2	Item Preequating.	219
6.9.3	Other Designs	221
6.10	Equating with Polytomous IRT	221
6.10.1	Polytomous IRT Models for Ordered Responses.	222
6.10.2	Scoring Function, Item Response Function, and Test Characteristic Curve.	227
6.10.3	Parameter Estimation and Scale Transformation with Polytomous IRT Models.	228

6.10.4	True Score Equating	232
6.10.5	Observed Score Equating.	232
6.10.6	Example Using the Graded Response Model	233
6.11	Robustness to Violations of the Unidimensionality Assumption	236
6.12	Practical Issues and Caveat	238
6.13	Exercises	239
	References	241
7	Standard Errors of Equating	247
7.1	Definition of Standard Error of Equating.	248
7.2	The Bootstrap	250
7.2.1	Standard Errors Using the Bootstrap	250
7.2.2	Standard Errors of Equating.	252
7.2.3	Parametric Bootstrap	253
7.2.4	Standard Errors of Equipercentile Equating with Smoothing	255
7.2.5	Standard Errors of Scale Scores	256
7.2.6	Standard Errors of Equating Chains	257
7.2.7	Mean Standard Error of Equating	258
7.2.8	Caveat.	259
7.3	The Delta Method	259
7.3.1	Mean Equating Using Single Group and Random Groups Designs	260
7.3.2	Linear Equating Using the Random Groups Design	261
7.3.3	Equipercentile Equating Using the Random Groups Design	263
7.3.4	Standard Errors for Other Designs	264
7.3.5	Illustrative Example	265
7.3.6	Approximations	267
7.3.7	Standard Errors for Scale Scores	268
7.3.8	Standard Errors of Equating Chains	269
7.3.9	Using Delta Method Standard Errors.	270
7.4	Using Standard Errors in Practice.	276
7.5	Exercises	278
	References	279
8	Practical Issues in Equating	283
8.1	Equating and the Test Development Process	285
8.1.1	Test Specifications	285
8.1.2	Changes in Test Specifications	286
8.1.3	Characteristics of Common-Item Sets	287

8.2	Data Collection: Design and Implementation	289
8.2.1	Choosing Among Equating Designs	289
8.2.2	Developing Equating Linkage Plans	292
8.2.3	Examinee Groups Used in Equating	300
8.2.4	Sample Size Requirements	303
8.3	Choosing from Among the Statistical Procedures	305
8.4	Equating Criteria and Designs in Research Studies	310
8.4.1	Criteria and Designs Based on Error in Estimating Equating Relationships	310
8.4.2	Equating in a Circle	318
8.4.3	Criteria and Designs Based on Assessing Group Invariance of Equating Relationships	319
8.4.4	Criteria and Designs Based on the Equity Property of Equating	320
8.4.5	Discussion of Equating Criteria and Designs	325
8.5	Choosing from Among Equating Results in Operational Equating	326
8.5.1	Equating Versus Not Equating	326
8.5.2	Use of Robustness Checks	327
8.5.3	Choosing from Among Results in the Random Groups Design	327
8.5.4	Choosing from Among Results in the Common-Item Nonequivalent Groups Design	328
8.5.5	Use of Consistency Checks	329
8.5.6	Equating and Score Scales	330
8.6	Importance of Standardization Conditions and Quality Control Procedures	331
8.6.1	Test Development	331
8.6.2	Test Administration and Standardization Conditions	331
8.6.3	Quality Control	333
8.6.4	Reequating	334
8.7	Conditions Conducive to Satisfactory Equating	337
8.8	Comparability Issues in Special Circumstances	337
8.8.1	Comparability Issues with Computer-Based Tests	337
8.8.2	Comparability for Constructed-Response and Mixed-Format Tests	344
8.8.3	Score Comparability with Optional Test Sections	348
8.9	Conclusion	349
8.10	Exercises	350
	References	352

9	Score Scales	371
9.1	Scaling Perspectives	372
9.2	Unit Scores, Item Scores, and Raw Scores.	377
9.2.1	Test Score Terminology	377
9.2.2	Unit and Item Scores	378
9.2.3	Raw Scores (Y)	380
9.3	Scores on Mixed-Format Tests	387
9.3.1	Weights Based on Numbers of Score Points	388
9.3.2	Observed Score Effective Weights	389
9.3.3	True Score Effective Weights.	390
9.3.4	Weights Chosen to Maximize Reliability.	390
9.3.5	Weighting Example.	391
9.3.6	Some Other Weighting Criteria and Issues.	392
9.3.7	Weights in IRT	392
9.4	Score Transformations.	393
9.5	Incorporating Normative Information	394
9.5.1	Linear Transformations	394
9.5.2	Nonlinear Transformations.	395
9.5.3	Example: Normalized Scale Scores.	397
9.5.4	Importance of Norm Group in Setting the Score Scale.	401
9.6	Incorporating Score Precision Information.	401
9.6.1	Rules of Thumb for Number of Distinct Score Points.	402
9.6.2	Linearly Transformed Score Scales with a Given Standard Error of Measurement	404
9.6.3	Score Scales with Approximately Equal Conditional Standard Errors of Measurement	405
9.6.4	Example: Incorporating Score Precision	407
9.6.5	Evaluating Psychometric Properties of Scale Scores.	410
9.6.6	The IRT θ -Scale as a Score Scale.	413
9.7	Incorporating Content Information	414
9.7.1	Item Mapping.	414
9.7.2	Scale Anchoring.	415
9.7.3	Standard Setting	417
9.7.4	Numerical Example.	418
9.7.5	Practical Usefulness	420
9.8	Maintaining Score Scales.	420
9.9	Scales for Test Batteries and Composites	422
9.9.1	Test Batteries	422
9.9.2	Composite Scores	423
9.9.3	Maintaining Scales for Batteries and Composites	424

9.10	Vertical Scaling and Developmental Score Scales	425
9.10.1	Structure of Batteries	427
9.10.2	Type of Domain Being Measured	428
9.10.3	Definition of Growth.	429
9.10.4	Designs for Data Collection for Vertical Scaling	431
9.10.5	Test Scoring.	434
9.10.6	Hieronimus Statistical Methods	435
9.10.7	Thurstone Statistical Methods.	437
9.10.8	IRT Statistical Methods	440
9.10.9	Thurstone Illustrative Example	445
9.10.10	IRT Illustrative Example	454
9.10.11	Statistics for Comparing Scaling Results	461
9.10.12	Some Limitations of Vertically Scaled Tests	463
9.10.13	Vertical Scaling Designs with Variable Sections.	465
9.10.14	Maintaining Vertical Scales	466
9.10.15	Research on Vertical Scaling	466
9.10.16	Score Scales and Growth Models	471
9.11	Exercises	473
	References	475
10	Linking	487
10.1	Linking Categorization Schemes and Criteria.	488
10.1.1	Types of Linking	491
10.1.2	Mislevy/Linn Taxonomy	492
10.1.3	Holland and Dorans Framework	496
10.1.4	Degrees of Similarity	498
10.1.5	Summary and Other Approaches	500
10.2	Group Invariance	501
10.2.1	Statistical Methods Using Observed Scores	501
10.2.2	Statistics for Overall Group Invariance	505
10.2.3	Statistics for Pairwise Group Invariance	507
10.2.4	Example: ACT and ITED Science Tests	508
10.3	Additional Examples.	527
10.3.1	Extended Time	528
10.3.2	Test Adaptations and Translated Tests.	529
10.4	Discussion	531
10.5	Exercises	532
	References	533
	Appendix A: Answers to Exercises.	537
	Appendix B: Computer Programs	559
	Index	561



<http://www.springer.com/978-1-4939-0316-0>

Test Equating, Scaling, and Linking
Methods and Practices

Kolen, M.J.; Brennan, R.L.

2014, XXVI, 566 p. 68 illus., Hardcover

ISBN: 978-1-4939-0316-0