

Contents

1. Introduction 1
2. Basic concepts 5
 - Searching the literature 6
 - Critical review 6
 - Empirical forms of validity 9
 - The two traditions of assessment 11
 - The reduction of measurement error 14
 - Summary 15
3. Devising the items 17
 - The source of items 18
 - Content validity 23
 - Generic versus specific scales and the 'fidelity versus bandwidth' issue 27
 - Translation 29
4. Scaling responses 37
 - Introduction 37
 - Some basic concepts 37
 - Categorical judgements 38
 - Continuous judgements 40
 - To rate or to rank 66
 - Multidimensional scaling 68
5. Selecting the items 77
 - Interpretability 77
 - Face validity 82
 - Frequency of endorsement and discrimination 83
 - Homogeneity of the items 85
 - Multifactor inventories 96
 - When homogeneity does not matter 97
 - Putting it all together 98
6. Biases in responding 103
 - The differing perspectives 103
 - Answering questions: the cognitive requirements 104
 - Optimizing and satisficing 108

Social desirability and faking good	110
Deviation and faking bad	115
Yea-saying or acquiescence	118
End-aversion, positive skew, and halo	119
Framing	122
Biases related to the measurement of change	123
Estimates of the prior state — implicit theory of change	125
Reconciling the two positions	125
Proxy reporting	126
Testing the items	127
7. From items to scales	135
Weighting the items	135
Missing items	139
Multiplicative composite scores	140
Transforming the final score	143
Percentiles	144
Standard and standardized scores	146
Normalized scores	149
Age and sex norms	149
Establishing cut points	151
Methods based on characteristics of the distribution	152
Methods based on judgement	154
Absolute methods	154
Receiver operating characteristic curves	156
Summary	163
8. Reliability	167
Basic concepts	167
Philosophical implications	170
Terminology	173
Defining reliability	174
Other considerations in calculating the reliability of a test	177
The observer nested within subject	179
Multiple observations	180
Other types of reliability	182
Different forms of the reliability coefficient	183
Kappa coefficient versus the ICC	188
The method of Bland and Altman	190
Issues of interpretation	190
Improving reliability	196

- Standard error of the reliability coefficient and sample size 198
- Reliability generalization 202
- The average value of r and α 203
- The variance of the reliability estimates 204
- Combining estimates 205
- Factors affecting the reliability 206
- Summary 207
- 9. Generalizability theory 211
 - Generalizability theory fundamentals 213
 - An Example 214
 - The First Step—the ANOVA 215
 - Step 2—From ANOVA to G coefficients 218
 - Relative vs. Absolute Error 219
 - Equivalent for the nested design 222
 - Generalizability of an average 222
 - Step 3 - from G study to D study 223
 - ANOVA for statisticians and ANOVA for psychometricians 224
 - Confidence intervals for G coefficients 225
 - The general rules to compute G coefficients 226
 - Getting the computer to do it for you 227
 - Some Common Examples 228
 - Uses and abuses of G theory 244
 - Summary 245
- 10. Validity 247
 - Why assess validity? 247
 - Reliability and validity 248
 - A history of the 'types' of validity 249
 - Content validation 252
 - Criterion validation 254
 - Construct validation 257
 - Construct validation studies 258
 - Extreme groups 261
 - Convergent and discriminant validation 262
 - Consequential validation 263
 - The multitrait–multimethod matrix 264
 - Summary 265
 - Responsiveness and sensitivity to change 266

Validity and 'types of indices'	267
Biases in validity assessment	268
Unreliability of the criterion	271
Changes in the sample	273
Validity generalization	274
Summary	274
11. Measuring change	277
Introduction	277
The goal of measurement of change	277
Why not measure change directly?	278
Measures of association—reliability and sensitivity to change	280
Difficulties with change scores in experimental designs	285
Change scores and quasi-experimental designs	286
Measuring change using multiple observations: growth curves	288
How much change is enough?	293
Summary	295
12. Item response theory	299
Problems with classical test theory	299
The introduction of item response theory	301
Item characteristic curves	302
The one-parameter model	304
The two- and three-parameter models	306
Polytomous models	309
Item information	312
Item fit	313
Person fit	315
Differential item functioning	315
Unidimensionality and local independence	316
The standard error of measurement	320
Equating tests	321
Sample size	322
Mokken scaling	323
Advantages	324
Disadvantages	326
Computer programs	327
13. Methods of administration	331
Face-to-face interviews	331
Advantages	331

Disadvantages	332
Telephone questionnaires	334
Random digit dialling	336
Advantages	337
Disadvantages	338
Mailed questionnaires	340
The necessity of persistence	346
Computer-assisted administration	348
Using e-mail and the Web	351
Personal data assistants	354
Reporting response rates	356
14. Ethical considerations	365
15. Reporting test results	373
Standards for Educational and Psychological Testing	374
The STARD initiative	376
Summary	379
Appendices	
A Further reading	381
B Where to find tests	387
C A (very) brief introduction to factor analysis	409
Author Index	415
Subject Index	423