

# Contents

- 1 Introduction** 1
  - References 3
- 2 Basic concepts** 4
  - Searching the literature 4
  - Critical review 5
  - The two traditions of assessment 9
  - Summary 12
  - References 12
- 3 Devising the items** 14
  - The source of items 15
  - Content validity 19
  - Generic versus specific scales and the 'fidelity versus bandwidth' issue 22
  - Translation 23
  - References 26
- 4 Scaling responses** 29
  - Introduction 29
  - Some basic concepts 29
  - Categorical judgements 30
  - Continuous judgements 32
  - To rate or to rank 53
  - Multidimensional scaling 54
  - References 56
- 5 Selecting the items** 61
  - Interpretability 61
  - Face validity 66
  - Frequency of endorsement and discrimination 67
  - Homogeneity of the items 68
  - Multifactor inventories 73
  - When homogeneity does not matter 74
  - Putting it all together 76
  - References 76

**6 Biases in responding 80**

- The differing perspectives 80
- Answering questions: the cognitive requirements 81
- Optimizing and satisficing 84
- Social desirability and faking good 85
- Deviation and faking bad 89
- Yea-saying or acquiescence 92
- End-aversion, positive skew, and halo 93
- Framing 95
- Biases related to the measurement of change 96
- References 98

**7 From items to scales 102**

- Weighting the items 102
- Multiplicative composite scores 105
- Transforming the final score 108
- Percentiles 109
- Standard and standardized scores 111
- Normalized scores 113
- Age and sex norms 113
- Establishing cut points 115
- Summary 123
- References 123

**8 Reliability 126**

- Basic concepts 126
- Philosophical implications 128
- Defining the reliability of a test 130
- Other considerations in calculating the reliability of a test 133
- Other types of reliability 137
- Different forms of the reliability coefficient 138
- Issues of interpretation 142
- Improving reliability 146
- Standard error of the reliability coefficient and sample size 148
- Summary 151
- References 151

**9 Generalizability theory 153**

G studies 155

D studies 155

Example 1—therapists, occasions, and patients 156

D study examples 160

Example 2—items, observers, and stations (the OSCE) 162

Example 3—econometric vs. psychometric perspectives on the utility of health states 164

Perspective 1: econometric 166

Perspective 2: psychometric 166

Perspective 3: experimental 166

General rules for generalizability 167

Nested designs 170

Error estimates for G coefficients 170

Summary 170

References 170

**10 Validity 172**

Why assess validity? 172

Reliability and validity 173

The ‘types’ of validity 174

Content validity 175

Criterion validity 176

Construct validity 178

Responsiveness and sensitivity to change 186

Validity and ‘types of indices’ 186

Biases in validity assessment 187

Changes in the sample 192

Summary 192

References 192

**11 Measuring change 194**

Introduction 194

The goal of measurement of change 194

Why not measure change directly? 195

Measures of association—reliability and sensitivity to change 196

Difficulties with change scores in experimental designs 201

Change scores and quasi-experimental designs 202

Measuring change using multiple observations: growth curves 204

How much change is enough? 209

Summary 210

References 210

**12 Item response theory 213**

- Item characteristic curves 214
- The one-parameter model 216
- The two- and three-parameter models 217
- Polytomous models 218
- Item fit 220
- Person fit 222
- The standard error of measurement 222
- Sample size 222
- Advantages 223
- Disadvantages 224
- Computer programs 225
- References 226

**13 Methods of administration 228**

- Face-to-face interviews 228
- Telephone questionnaires 231
- Mailed questionnaires 234
- The necessity of persistence 239
- Computer-assisted administration 241
- Using e-mail and the Web 243
- References 244

**14 Ethical considerations 248**

- References 253

**Appendices**

- A Further reading 254
- B Where to find tests 257
- C A (very) brief introduction to factor analysis 265

**Author Index 271**

**Subject Index 277**