## CONTENTS

Acknowledgments xv
A Foreword (in Berkian Style) by Mike Theall xix

### Introduction 1

1 TOP 13 SOURCES OF EVIDENCE OF TEACHING EFFECTIVENESS 9

- A Few Ground Rules 10
- Teaching Effectiveness: Defining the Construct 11
- National Standards 12
- Beyond Student Ratings 13
- A Unified Conceptualization 13

### Thirteen Sources of Evidence 14

- Student Ratings 15
- Peer Ratings 19
- External Expert Ratings 22
- Self-Ratings 23
- Videos 24
- Student Interviews 25
- Exit and Alumni Ratings 27
- Employer Ratings 28
- Administrator Ratings 29
- Teaching Scholarship 30
- Teaching Awards 31
- Learning Outcome Measures 32
THIRTEEN STRATEGIES TO MEASURE COLLEGE TEACHING

Teaching Portfolio 34
BONUS: 360° Multisource Assessment 37
Berks Top Picks 44
Formative Decisions 45
Summative Decisions 45
Program Decisions 45
Decision Time 45

CREATING THE RATING SCALE STRUCTURE 47

Overview of the Scale Construction Process 48
Specifying the Purpose of the Scale 48
Delimiting What Is to Be Measured 50
Focus Groups 50
Interviews 50
Research Evidence 51
Determining How to Measure the "What" 54
Existing Scales 55
Item Banks 55
Commercially Published Student Rating Scales 56
Universe of Items 58
Structure of Rating Scale Items 60
Structured Items 60
Unstructured Items 62

GENERATING THE STATEMENTS 65

Preliminary Decisions 66
Domain Specifications 66
Number of Statements 66
Rules for Writing Statements 66
1. The statement should be clear and direct. 69
2. The statement should be brief and concise. 69
3. The statement should contain only one complete behavior, thought, or concept. 71
4. The statement should be a simple sentence. 72
5. The statement should be at the appropriate reading level. 72
6. The statement should be grammatically correct. 73
7. The statement should be worded strongly. 73
8. The statement should be congruent with the behavior it is intended to measure. 74
9. The statement should accurately measure a positive or negative behavior. 74
10. The statement should be applicable to all respondents. 75
11. The respondents should be in the best position to respond to the statement. 76
12. The statement should be interpretable in only one way. 78
13. The statement should NOT contain a double negative. 78
14. The statement should NOT contain universal or absolute terms. 79
15. The statement should NOT contain nonabsolute, warm-and-fuzzy terms. 79
16. The statement should NOT contain value-laden or inflammatory words. 80
17. The statement should NOT contain words, phrases, or abbreviations that would be unfamiliar to all respondents. 81
18. The statement should NOT tap a behavior appearing in any other statement. 81
19. The statement should NOT be factual or capable of being interpreted as factual. 82
20. The statement should NOT be endorsed or given one answer by almost all respondents or by almost none. 83

4 SELECTING THE ANCHORS 85

Types of Anchors
Intensity Anchors 86
Evaluation Anchors 87
Frequency Anchors 89
Quantity Anchors 90
Comparison Anchors 91

Rules for Selecting Anchors 94
1. The anchors should be consistent with the purpose of the rating scale. 94
2. The anchors should match the statements, phrases, or word topics. 96
3. The anchors should be logically appropriate with each statement. 98
4. The anchors should be grammatically consistent with each question. 99
5. The anchors should provide the most accurate and concrete responses possible. 100
6. The anchors should elicit a range of responses.
7. The anchors on bipolar scales should be balanced, not biased.
8. The anchors on unipolar scales should be graduated appropriately.

5 REFINING THE ITEM STRUCTURE

Preparing for Structural Changes

1. What rating scale format is best?
2. How many anchor points should be on the scale?
3. Should there be a designated midpoint position, such as “Neutral,” “Uncertain,” or “Undecided,” on the scale?
4. How many anchors should be specified on the scale?
5. Should numbers be placed on the anchor scale?
6. Should a “Not Applicable” (NA) or “Not Observed” (NO) option be provided?
7. How can response set biases be minimized?
CONTENTS

8 COLLECTING EVIDENCE OF VALIDITY AND RELIABILITY 161

Validity Evidence 162
- Evidence Based on Job Content Domain 164
- Evidence Based on Response Processes 168
- Evidence Based on Internal Scale Structure 169
- Evidence Related to Other Measures of Teaching Effectiveness 171
- Evidence Based on the Consequences of Ratings 172

Reliability Evidence 174
- Classical Reliability Theory 174
- Summated Rating Scale Theory 175
- Methods for Estimating Reliability 176

Epilogue 182

9 REPORTING AND INTERPRETING SCALE RESULTS 185

Generic Levels of Score Reporting 186
- Item Anchor 186
- Item 187
- Subscale 190
- Total Scale 191
- Department/Program Norms 192
- Subject Matter/Program-Level State, Regional, and National Norms 193

Criterion-Referenced versus Norm-Referenced Score Interpretations 194
- Score Range 194
- Criterion-Referenced Interpretations 195
- Norm-Referenced Interpretations 196

Formative, Summative, and Program Decisions 197
- Formative Decisions 198
- Summative Decisions 204
- Program Decisions 212

Conclusions 213

References 215

Appendices 241
- A. Sample “Home-Grown” Rating Scales 241
- B. Sample 360° Assessment Rating Scales 257
- C. Sample Reporting Formats 273
- D. Commercially Published Student Rating Scale Systems 277

Index 281