Contents

Acknowledgements x
Preface xi

1 *Introduction: the rubber ruler* 1
1.1 Why test language learning? 1
1.2 What is a language test? 2
1.2.1 What are the properties of measuring devices? 2
1.3 The rubber ruler 3
1.4 Tests, measurement and evaluation 5
1.4.1 Evaluation without measurement 5
1.4.2 Measurement without a test 5
1.4.3 Tests 6
1.5 Ethical uses of language tests 9
1.5.1 Reliability 10
1.5.2 Validity 10
Further reading 13
Exercises 13

2 Measuring language ability and making decisions 17
2.1 Measuring language ability 17
2.2 Making decisions about learners, teachers, programmes, and policies 19
2.3 Contexts of language use 20
2.3.1 Contextual features 21
2.3.2 Authenticity 24
2.4 Making valid interpretations of test performance 26
2.4.1 Consistency of measurement 26
2.4.2 Validity: evidence for interpretations of test performance 29
2.5 Conclusion: bias for best 34
Further reading 35
Exercises 35

3 Test development 38
3.1 What do I need to test? Needs analysis 38
3.1.1 Define the purpose of the test 40
3.1.2 Conduct a preliminary investigation 40

3.1.3 Collect primary data 40
3.1.4 Collect secondary data 42
3.1.5 Analyse target language use task and language characteristics 44
3.2 How am I going to test language abilities? Turning target language use tasks into test tasks 48
3.2.1 Developing a test task 48
3.2.2 Developing a blueprint for the test 49
3.2.3 Options for test tasks 49
3.3 How am I going to give the test? Test administration 54
3.3.1 Test environment 54
3.3.2 Personnel 54
3.3.3 Procedures 55
3.3.4 Scoring 55
3.4 How can my computer assist me in test development? Computer-based tools 56
3.4.1 *Hot Potatoes* 57
3.4.2 *Moodle* 60
3.4.3 WebCT 60
3.5 Conclusion 63
Further reading 64
Exercises 64

4 Alternatives in assessment 67
4.1 Norm-referenced and criterion-referenced tests 67
4.2 Communicative language tests 69
4.3 'General' and 'specific purpose' language tests 70
4.4 Discrete-point and integrative tests 70
4.5 Formative and summative assessment 72
4.6 Alternative approaches to assessment 73
4.6.1 Conference assessments 74
4.6.2 Portfolio assessment 74
4.6.3 Self- and peer-assessments 75
4.6.4 Task-based and performance assessment 76
4.6.5 Dynamic assessment 79
4.6.6 Summary 80
4.7 Conclusion 80
Further reading 81
Exercises 81

5 By the numbers: a statistics mini-course 85
5.1 Introduction 85
5.2 Normal distribution 87
5.3 The average or mean 88
5.4 Standard deviation 90

5.4.1 Standard deviation as a unit of measurement 92
5.5 Correlation 93
5.6 Probability and statistical significance 97
5.7 The t-test of the difference between two averages 99
5.8 Analysis of variance 101
5.9 Reliability 104
5.9.1 Split-half method 105
5.9.2 Internal consistency method 106
5.9.3 Standard error of measurement 108
5.10 The reliability of human raters 110
5.11 Conclusion 111
Further reading 112
Exercises 113

6 Technology and language testing 115
6.1 Introduction 115
6.2 Issues in technology and language testing 116
6.2.1 Technology and test taker attitudes 117
6.2.2 Language performance and different media 117
6.2.3 Technology and the construct to be measured 118
6.2.4 Technology and assessment tasks 118
6.2.5 The limits of automated scoring 119
6.3 Technology and language task types 119
6.3.1 Listening tasks 119
6.3.2 Integrated listening and speaking tasks 122
6.3.3 Writing tasks 124
6.3.4 Reading tasks 125
6.4 The promise and threats of automated scoring 127
6.4.1 Examples of current automated scoring programs 127
6.4.2 Concerns about automated scoring 129
6.5 Test feedback and reporting 131
6.6 Online and computer-based resources for statistics 135
6.6.1 Microsoft® Excel 135
6.6.2 Online resources 137
6.7 Conclusion 139
Further reading 139
Exercises 140

Afterword: the rubber ruler revisited 144
References 146
Index 153