Table of Contents

1 INTRODUCTION 11

1.1 Purpose statement 12
1.2 Structure of the text 12

2 FUNDAMENTAL CONCEPTS IN LANGUAGE TESTING 15

2.1 A Brief Historical Context 15
2.1.1 Pre-scientific era 15
2.1.2 Psychometric-structuralist era 16
2.1.3 Psycholinguistic-socio-linguistic era 17
2.1.4 Communicative language testing 18
2.1.5 Formative testing 19
2.1.6 Assessment for learning 20
2.2 Test Qualities 21
2.2.1 Validity 21
2.2.1.1 Content validity 22
2.2.1.2 Face validity 23
2.2.1.3 Response validity 24
2.2.1.4 Criterion-related validity aspects 24
2.2.1.5 Construct validity 25
2.2.2 Reliability 26
2.2.3 Validity and Reliability tensions 27
2.2.4 Practicality 27
2.2.5 Washback 29
2.2.6 Providing Feedback on test performance 30
2.2.7 Practicality and Washback tensions 31
2.3 Assessment Dichotomies 32
2.3.1 Criterion-referenced and Norm-referenced tests 32
2.3.2 Discrete-point versus Integrative tests 33
2.3.3 Summative versus Formative assessment 34
2.3.4 Competence versus Performance 36

3 COMPUTER-ASSISTED LANGUAGE TESTING 39

3.1 Digital Literacy 39
3.2 Towards a CALT Framework 42
3.2.1 Directionality 44
3.2.2 Delivery Format 46
3.2.3 Media Density 46
3.2.4 Target Skill 47
3.2.5 Scoring Mechanisms 48
3.2.6 Stakes 49
3.2.7 Purpose 50
3.2.8 Response Type 52
3.2.9 Task Types 52

3.2.9.1 Selected Response Tasks 54
3.2.9.2 Limited Production tasks 56
3.2.9.3 Effective Item Checklist 57
3.3 CALT Advantages and Disadvantages 59
3.3.1 CALT Validity Threats 65
3.4 CALT Innovative Task Types 67
3.5 Comparability Studies 69
3.5.1 Individual Differences 71

4 RESEARCH METHODOLOGY 75

4.1 Research Approach 75
4.2 Research Questions 76
4.3 Participants 77
4.4 Research Apparatus 78
4.5 Research and Data Collection Procedures 80
4.6 Limitations 83

5 DATA ANALYSIS 85

5.1 Pilot Quantitative Data Analysis 86
5.1.1 Descriptives, Limitations and Discussion 86
5.2 Study 1 Quantitative Data Analysis 90
5.2.1 Reliability Estimates 90
5.2.2 Descriptives 92
5.2.3 Comparing Mean Test Scores 94
5.2.4 Study 1 Discussion, Limitations and Conclusions 99
5.3 Study 2 Quantitative Data Analysis 101
5.3.1 Reliability Estimates 101
5.3.2 Descriptives 103
5.3.3 Comparing Mean Test Scores 106
5.3.4 Comparing Item Means 109
5.3.5 Study 2 Discussion, Limitations and Conclusions 111
5.4 Pilot Qualitative Data Analysis 113
5.4.1 Positives of the Pencil and Paper-based test 113
5.4.2 Negatives of the Pencil and Paper-based test 113
5.4.3 Positives of the Computer-based test 114
5.4.4 Negatives of the Computer-based test 114
5.5 Study 1 Qualitative Data Analysis 115
5.5.1 Positives of the Pencil and Paper-based test 115
5.5.2 Negatives of the Pencil and Paper-based test 117
5.5.3 Positives of the Computer-based test 119
5.5.4 Negatives of the Computer-based test 120
5.6 Study 2 Qualitative Data Analysis 122
5.6.1 Positives of the Pencil and Paper-based test 123
5.6.2 Negatives of the Pencil and Paper-based test 125
5.6.3 Positives of the Computer-based test 127
5.6.4 Negatives of the Computer-based test 128
5.7 Research Tool Development 131

6 MERGING THE DATA: FINDINGS AND DISCUSSION 135

6.1 Research Question 1 135
6.2 Research Question 2 136
6.3 Research Question 3 136
6.4 Research Question 4 137
6.5 Research Question 5 140
6.6 Research Question 6 143
6.7 Practical Implications 144

7 CONCLUSION 147

7.1 Further Research 150
7.2 Closing Statement 151

REFERENCES 153

APPENDICES 165

Appendix 0: A link to a shared storage space on a drive 165
Appendix 1: Quantitative data analysis PILOT 165
Appendix 2: Quantitative Data Analysis Study 1 166
Appendix 3: Quantitative Data Analysis Study 2 168
Appendix 4: Qualitative Data Analysis 170
Appendix 5: Statistical Model – Preferences 173
Model: Vocabulary subtest (mean subtest scores) 173
Model: Vocabulary subtest (mean subtest scores) 175