

**What Is the Relationship Between Alignment and Washback?**  
**A Mixed-Methods Study of the Libyan EFL Context**

By

Nwara Abdulhamid

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of  
the requirements for the degree of  
Doctor of Philosophy

Carleton University  
Ottawa, Canada

Copyright© Nwara Abdulhamid, 2018

## Abstract

This research uses an explanatory sequential mixed-methods approach (Creswell, 2015) to examine the relationship between the degree of alignment of components of the Libyan education system and the *washback* of the revised Secondary Education Certificate Examination of English (rSECEE). Washback is viewed as the influence of tests on both teaching and learning within the classroom. Some have argued (e.g., Linn, 2000; Shohamy, 1997; Tan, 2008) that lack of alignment between components of an education system may result in negative washback. Applying quantitative methods, Phase I of the study, draws on Webb's (1997) alignment model to investigate the degree of alignment between the rSECEE, Libya's English as a Foreign Language (EFL) content standards, and curriculum (i.e. the standardised textbooks in this Libyan context). Subsequently, applying qualitative methods, Phase II elicited data from three Libyan EFL teachers and their students through questionnaires, semi-structured interviews, classroom observations, and focus group interviews. Both quantitative and qualitative data were analysed, synthesized and merged into one overall interpretation of the phenomenon (Creswell, 2000). The Phase I and II results indicate that: 1) there is limited-to-no degree of alignment between the rSECEE and the EFL content standards; 2) the rSECEE does not meet Webb's (1997) comprehensive criteria; 3) rSECEE appears to have had negative washback on some teachers and their teaching, but little-to-no negative washback on other teachers; 4) rSECEE may have had negative washback on learners and their learning. Accordingly, it can be argued that the washback of rSECEE is highly complex; how teachers react is highly individual; but, teachers are the key stake-holders in mediating washback. The test design and its degree of alignment with the focal construct, the prevailing *culture of learning* (Cortazzi & Jin, 1992), and teachers' beliefs (Woods, 1996) seems to determine the *direction, intensity* (Cheng, 2005; Green, 2007,

Watanabe, 1996), and *magnitude* of washback (Fox, 2018). This study will be of interest to test developers, administrators, and policy makers, as these stake-holders require comprehensive assessment instruments that provide valid inferences about students' achievements, without undermining learning opportunities.

## Acknowledgements

I thank Allah, the Almighty for His greatness and for giving me the strength, courage, and perseverance to complete this journey. Without Allah's blessing this work would have not been possible.

This work would not have been possible without the help and support of many individuals. I wish to acknowledge and express my gratitude and thanks to these people for their valuable and countless contributions.

I owe my sincere gratitude to my supervisor Professor Janna Fox, for all the time and expertise given during the course of this study. Her expertise in the field, supervision, support and patience during both my Masters and Doctoral programmes have enabled me to develop a much more thorough understanding of the subject at hand. Her constant encouragement during the final days of this journey kept me motivated. She constantly wrote in her e-mails to me "*Nwara! Keep it up, it will pay off in the end, Hugs Janna*".

I am also grateful to my committee members, Professor Pychyl and Professor Cheng. I would like to thank them for their care and interest in my work. Their encouragement, advice and constructive feedback helped me in many ways to revise my work.

I would like to give thanks to my beloved *family* for being my source of strength and inspiration during this journey. My *Parents*, who believed in me and made me the person I am today. My *father*, who was the source of inspiration that guided me to further my education. He constantly calmed my panicking moments by saying over the phone "*Everything will be fine, and it will soon be a story for you to tell*". My *mother's* affection, love, and prayers that I felt in the air despite the long distance between us nurtured this work. My thanks also to my four brothers, in particular Abdulhamid for his efforts to gather all the necessary documents that I

needed to complete this work. My gratitude coupled with tears of happiness go to my only sister, *Asia*. Thank you my dear for all those hours of patience and constant encouragement. Thank you for listening to me cry and providing me with a boost of energy whenever I needed it.

My thanks also go to my friends, in particular Hajer Algezeeri, who never hesitated to offer a hand and support with the data collection. My honest appreciation is also extended to the two school principals and the administrative staff for their sincere time and cooperation. I am also grateful to the test-developer, three Grade 12 teachers and their students who participated in the study. I am truly grateful for their time, effort, and commitment.

Finally, and above all, I am grateful to my husband, *Jamal* for all his sacrifices and patience. Many thanks coupled with hugs and tears go to my four children Jihad, Maria, Myre and Owais. Thank you, children for being understanding and doing all the chores on my behalf. Thank you, Maria and Myre for pushing me to go to the library and assuring me you that would be responsible. I would have not made it without Jamal's and my children's' unwavering love and support.

## Table of Content

|  |      |
|--|------|
| <b>Abstract</b> .....  | ii   |
| <b>Acknowledgements</b> .....  | iv   |
| <b>List of Tables</b> .....  | xii  |
| <b>List of Figures</b> .....   | xiv  |
| <b>List of Abbreviations</b> .....   | xvi  |
| <b>Glossary of Terms</b> .....   | xvii |
| <b>Chapter I</b> .....   | 1    |
| <b>Introduction</b> .....  | 1    |
| <b>1.1. Background</b> .....   | 1    |
| <b>1.2. Purpose of the Study</b> .....   | 6    |
| <b>1.3. Research Questions</b> .....   | 9    |
| <b>1.4. Overview of the Thesis</b> .....   | 9    |
| <b>CHAPTER II</b> .....  | 11   |
| <b>Background</b> .....  | 11   |
| <b>2.1. Historic Background</b> .....  | 11   |
| <b>2.2. Languages</b> .....  | 14   |
| <b>2.3. The Evolution of Libyan Education</b> .....  | 15   |
| <b>2.3.1. Development of Education before Independence</b> .....                                       | 15   |
| <b>2.3.2. Development of Education since Independence</b> .....  | 16   |
| <b>2.4. English Language Teaching in Libya</b> .....   | 19   |
| <b>2.5. The Secondary School System and Its Development: 1960-2017</b> .....                           | 27   |
| 2.5.2. Schools and Classes at the Secondary Level.....   | 31   |
| 2.4.2. Characteristics of the Educational and Classroom Culture.....                                   | 32   |
| 2.4.3. Secondary Level EFL Teachers .....  | 37   |
| 2.4.4. Challenges of Teaching English at the Secondary Level.....                                      | 39   |
| <b>2.6. The Role of Examinations in Libya</b> .....  | 43   |
| 2.6.2. The SECEE .....   | 48   |
| <b>Chapter III</b> .....   | 51   |
| <b>Literature Review</b> .....   | 51   |
| <b>3.1. The Relationship Between Standards, Testing, and Instructional Activities and Materials</b> .. | 51   |
| 3.1.1. Standards.....  | 51   |
| 3.1.2. Assessment.....   | 52   |
| 3.1.3. Teaching.....   | 55   |

|   |            |
|---|------------|
| 3.1.4. The Triangular Relationship between Standards, Testing, and Teaching and Learning..... | 55         |
| <b>3.2. Overview of Alignment.....</b>  | <b>59</b>  |
| 3.2.1. Extended Review of Alignment.....  | 60         |
| 3.2.2. A Contemporary Definition of Validity.....   | 63         |
| 3.2.3. Alignment Research Models.....   | 71         |
| 3.2.3.1. The Webb Model.....  | 74         |
| 3.2.3.2. SEC model.....   | 80         |
| 3.2.3.3. Achieve model.....   | 80         |
| 3.2.3.4. Commonalities and differences across the three approaches of alignment.....          | 83         |
| <b>Chapter IV.....</b>  | <b>86</b>  |
| <b>Washback and High-Stakes Testing.....</b>  | <b>86</b>  |
| <b>4.1. Washback.....</b>   | <b>86</b>  |
| 4.1.1. Definition of Washback.....  | 86         |
| 4.1.2. Washback: Positive, Negative, Neutral or Both.....                                     | 88         |
| 4.1.2.1. Negative Washback on Teachers and Teaching.....                                      | 90         |
| 4.1.2.2. Washback on Learners and Learning.....   | 94         |
| 4.1.2.3. Positive Washback on Teachers and Learners.....                                      | 96         |
| <b>4.2. Overview of Washback.....</b>   | <b>99</b>  |
| <b>4.3. The Washback Trends in Language Testing Literature.....</b>                           | <b>100</b> |
| <b>4.4. Washback Models.....</b>  | <b>105</b> |
| Hughes (1993).....  | 105        |
| Bailey (1996).....  | 106        |
| Alderson and Hamp-Lyons (1996).....   | 107        |
| Burrows (2004).....   | 108        |
| Cheng (2005).....   | 112        |
| Green (2007).....   | 112        |
| <b>Summary.....</b>   | <b>114</b> |
| <b>Chapter V.....</b>   | <b>118</b> |
| <b>Methodology.....</b>   | <b>118</b> |
| <b>5.1. Research Questions.....</b>   | <b>118</b> |
| <b>5.2. Presuppositions of the rSECEE Operation.....</b>                                      | <b>120</b> |
| Phase I.....  | 120        |
| Phase II.....   | 120        |
| <b>5.3. Research Philosophy.....</b>  | <b>121</b> |
| <b>5.4. Research Design.....</b>  | <b>123</b> |

|  |            |
|--|------------|
| 5.4.1. Justification of the Research Design .....  | 124        |
| <b>5.5. Methods</b> .....  | 127        |
| <b>5.5.1. Participants</b> .....   | <b>128</b> |
| <b>5.6. Sampling</b> .....   | 130        |
| 5.6.1. Phase I.....  | 130        |
| 5.6.2. Phase II.....   | 130        |
| <b>5.7. Data Collection</b> .....  | 130        |
| 5.7.1. Phase I.....  | 131        |
| 5.7.1.1. Document Analysis .....   | 131        |
| 5.7.1.2. Alignment Analysis Process.....   | 132        |
| 5.7.2. Phase II.....   | 132        |
| 5.7.2.1. Instruments .....   | 133        |
| <b>5.8. Analysis</b> .....   | 139        |
| <b>5.9. Reliability and Trustworthiness</b> .....  | 139        |
| <b>CHAPTER VII</b> .....   | 142        |
| <b>Phase I</b> .....   | 142        |
| <b>6.1. Method</b> .....   | 143        |
| 6.1.1 Participants.....  | 143        |
| 6.1.2. Policy-Makers .....   | 144        |
| <b>6.2. Data Collection</b> .....  | 144        |
| 6.2.2. Document Analysis .....   | 144        |
| 6.2.3. Alignment Analysis Process .....  | 145        |
| 6.2.4. Instruments.....  | 146        |
| 6.2.4.1. Libya’s EFL Academic Content Standards for Teaching English in Libyan Secondary<br>Classrooms ..... | 146        |
| 6.2.4.2. The 2014/2015 rSECEE .....  | 147        |
| 6.2.4.3. Semi-Structured Interviews.....   | 148        |
| <b>6.2.5. Procedure</b> .....  | <b>148</b> |
| 6.2.5.1. Identifying Criteria and Acceptable Levels.....   | 148        |
| 6.2.5.2. Developing the Coding Matrix for the Content Domain.....  | 149        |
| 6.2.5.3. Training Alignment Review Panel Members .....   | 150        |
| 6.2.5.4. The Alignment Review Panel Coding.....  | 151        |
| <b>6.3. Analysis</b> .....   | 155        |
| 6.3.2. Document Analysis .....   | 155        |
| 6.3.3. Alignment Analysis Process .....  | 156        |

|   |     |
|---|-----|
| 6.3.3.1. Categorical Concurrence .....  | 157 |
| 6.3.3.2. Depth-of-Knowledge (DOK) Consistency .....                               | 158 |
| 6.3.3.3. Range-of-Knowledge Correspondence .....                                  | 159 |
| 6.3.3.4. The Balance of Representation.....                                       | 159 |
| <b>6.4. Results</b> .....   | 161 |
| 6.4.2. English-Secondary EFL Textbooks .....                                      | 161 |
| 6.4.3. The rSECEE.....  | 170 |
| 6.4.4. Webb Model (1997, 1999) Alignment Results .....                            | 173 |
| 6.4.4.1. Categorical Concurrence .....  | 173 |
| 6.4.4.2. Depth-of-Knowledge (DOK) Consistency .....                               | 174 |
| 6.4.4.3. Range-of-Knowledge Correspondence .....                                  | 176 |
| 6.4.4.4. The Balance of Representation.....                                       | 177 |
| 6.4.4.5. Panel Review Responses .....   | 178 |
| 6.4.5. Test-Developer’s Responses .....   | 182 |
| <b>6.5. Discussion</b> .....  | 185 |
| <b>Chapter VII</b> .....  | 200 |
| <b>Phase II Study</b> .....   | 200 |
| <b>7.1 Research Questions</b> .....   | 201 |
| <b>7.2 Method</b> .....   | 202 |
| 7.2.1 Research Setting .....  | 202 |
| 7.2.1.1 Description of High Schools UM and AL.....                                | 202 |
| 7.2.2 Participants .....  | 204 |
| 7.2.2.1 Teachers and Students.....  | 204 |
| <b>7.3 Data Collection</b> .....  | 206 |
| 7.3.1 Instruments .....   | 206 |
| 7.3.1.1 Stage 1: Questionnaires.....  | 208 |
| 7.3.1.2 Stage 2: Observation .....  | 208 |
| 7.3.1.3 Stage 3: Semi-Structured Interviews.....                                  | 211 |
| 7.3.1.4. Stage 4: Focus Groups .....  | 213 |
| <b>7.4 Analysis</b> .....   | 215 |
| 7.4.1 Questionnaires .....  | 216 |
| 7.4.2 Observations .....  | 217 |
| 7.4.3 Semi Structured and Focus Group Interviews with Teachers and Students ..... | 218 |
| <b>7.5. Results</b> .....   | 220 |
| 7.5.1. Washback on Teachers .....   | 220 |

|   |            |
|---|------------|
| 7.5.1.1 The Revised Curriculum .....  | 220        |
| 7.5.1.2 Teachers' Feelings .....  | 222        |
| 7.5.1.3 Impact of the rSECEE on the Content of the Teaching .....   | 227        |
| 7.5.1.4 Impact of rSECEE on Methods of Teaching. ....   | 231        |
| 7.5.1.5 Impact of rSECEE on Learning. ....  | 243        |
| 7.5.1.6 Impact of rSECEE on Classroom Assessment.....   | 246        |
| 7.5.1.7. Impact of rSECEE on Students.....  | 248        |
| 7.5.2 Washback on Students.....   | 251        |
| 7.5.2.1 Students' Attitudes Towards the rSECEE.....   | 251        |
| 7.5.2.2 Students' Attitudes Towards the rSECEE and Grade 12 EFL Teaching .....  | 257        |
| <b>4. 7.6. Discussion.....</b>  | <b>260</b> |
| 7.6.1 The Washback Effect on Teachers and Teaching.....   | 262        |
| 7.6.1.1 The Washback Effect on Teachers.....  | 262        |
| 7.6.1.2. The Washback Effect on Teaching Content.....   | 265        |
| 7.6.1.3 The Washback Effect on Instructional Practices .....  | 268        |
| 7.6.1.4. Burrows' (2004) Framework of Analysis.....   | 274        |
| 7.6.2. Washback on Learners and Learning.....   | 275        |
| 7.6.2.1 The Washback Effect on Learners .....   | 275        |
| 7.6.2.2 The Washback Effect on Student Learning .....   | 277        |
| <b>7.7 Merging the Phase I and Phase II Findings.....</b>   | <b>281</b> |
| 7.7.1 The Lack of Alignment of the rSECEE and Washback.....   | 283        |
| 7.7.2 The Libyan Culture of Learning Factor.....  | 290        |
| 7.7.3 The Teacher Factor .....  | 293        |
| <b>Chapter XIII.....</b>  | <b>297</b> |
| <b>Conclusion .....</b>   | <b>297</b> |
| <b>8.1. Summary of Study .....</b>  | <b>297</b> |
| <b>8.2. Implications for Key Stake-holders in Libya.....</b>  | <b>301</b> |
| <b>8.3. Recommendations .....</b>   | <b>302</b> |
| <b>8.4 Limitations and Future Research.....</b>   | <b>308</b> |
| <b>References.....</b>  | <b>311</b> |
| <b>Appendix A: Former Version of BECEE which is similar to the former SECEE .....</b>   | <b>368</b> |
| <b>Appendix B: A Sample Test of the rSECEE.....</b>   | <b>372</b> |
| <b>Appendix C: Review on Previous and Current Washback Studies .....</b>  | <b>381</b> |
| <b>Appendix D: Summary of Factors Identified by Empirical Studies as Affecting the Degree and Kinds of Washback (Source, Spratte, 2005, p. 29).....</b> | <b>386</b> |

|   |     |
|---|-----|
| <b>Appendix E: Teacher Questionnaire</b> .....  | 387 |
| <b>Appendix F: Ethics Clearance Letter</b> .....  | 390 |
| <b>Appendix G: Student Questionnaire</b> .....  | 392 |
| <b>Appendix H: Teacher Interview Questions</b> .....  | 394 |
| <b>Appendix I: Grade 12 Content Objectives</b> .....  | 395 |
| <b>Appendix J: Policy Maker Interview Questions</b> .....   | 403 |
| <b>Appendix K: Test Developer Interview Questions</b> .....   | 404 |
| <b>Appendix L: World Language Cognitive Rigor Matrix (CRM, 2015) for the EFL descriptions of the depth of knowledge (DOK)</b> ..... | 405 |
| <b>Appendix M: Coding Matrix I</b> .....  | 406 |
| <b>Appendix N: Coding Matrix II</b> .....   | 415 |
| <b>Appendix O: A Sample Unit of the Grade 12 EFL Literary Course Textbook</b> .....   | 416 |
| <b>Appendix P: Student Focus Group Questions</b> .....  | 423 |
| <b>Appendix Q: Extra Detail on Data Collection Procedures</b> .....   | 424 |
| <b>Appendix R: An Example of an Observational Coding Sheet</b> .....  | 427 |
| <b>Appendix S: Samples of Teachers' Participants classroom assessment</b> .....   | 428 |
| Teacher A.....  | 428 |
| Teacher B.....  | 428 |
| Teacher C.....  | 429 |
| <b>Appendix T: Samples of Teachers C's white board work</b> .....   | 432 |

## List of Tables

|  |     |
|--|-----|
| Table 2.1: Secondary Schools Majors.   | 29  |
| Table 2.2: Results of the Secondary Education Certificate Examination of English Language Specialization Major, 2006-2014. | 30  |
| Table 2.3: Differences between the rSECEE and the Original SECEE   | 47  |
| Table 2.4: Results of the SECE, 2015 – 2017 Scientific Section   | 50  |
| Table 2.5: Results of the SECE, 2015 – 2017 Literacy Section   | 50  |
| Table 3.1: Messick’s Facets of Validity  | 65  |
| Table 3.2: Overview of Major Alignment Models  | 73  |
| Table 3.3: The Complete Set of Webb’s 1997 Alignment Criteria  | 75  |
| Table 3.4: Webb’s General Descriptions for Depth-Of-Knowledge Levels   | 77  |
| Table 3.5: Summary of the Three Alignment Models   | 82  |
| Table 4.1: A Review of The Current Washback Trends   | 104 |
| Table 5.1: Reasons for Using Mixed Methods Design and Approaches   | 126 |
| Table 6.1: Chronological Distribution of Lessons within the Revised Grade 12 EFL Literary and Scientific Course Textbooks  | 163 |
| Table 6.2: Examples of the Libyan EFL Curricular Standard Being Operationalised within the Assigned Textbooks              | 168 |
| Table 6.3: Examples of Poorly Crafted Test Items in the 2014/2015 rSECEE   | 172 |
| Table 6.4: Reviewers’ Agreement on Coding  | 173 |
| Table 6.5: Categorical Concurrence of the rSECEE   | 174 |
| Table 6.6: A Comparison of DOK Target Levels to the rSECEE DOK Levels  | 174 |
| Table 6.7: DOK Consistency for the rSECEE  | 175 |

|   |     |
|---|-----|
| Table 6.8: Range of Knowledge Criterion for the rSECEE                                    | 176 |
| Table 6.9: Balance of Representation Criterion for the rSECEE                             | 178 |
| Table 6.10: Examples of the rSECEE Test Items Measuring Grade 12 Students' Reading Skills | 194 |
| Table 7.1: Numbers of Teachers in UM and AL High Schools                                  | 203 |
| Table 7.2: AL High School Students Pass Rates in rSECEE from 2015-2017                    | 203 |
| Table 7.3: UM High School Students Pass Rates in rSECEE from 2015-2017                    | 203 |
| Table 7.4: Teacher Demographics and Classroom Characteristics                             | 205 |
| Table 7.5: Extent of Stated Changes for Participant Teachers                              | 275 |

**\*Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes.**

## List of Figures

|  |     |
|--|-----|
| Figure 2.1: Structure of the Formal Education System in Libya  | 19  |
| Figure 2.2: Typical Classroom Arrangement in a Libyan Secondary School   | 32  |
| Figure 3.1: Relationship among standards, instructional activities, and assessment/tests   | 56  |
| Figure 3.2: Alignment as Links between the components of an Education System   | 61  |
| Figure 3.3: The Role of Test Specifications in the Stages of Test Development  | 67  |
| Figure 3.4: Case 1: The Relationship among Standards, Assessment Specifications and Assessment Regarding Validity and Alignment Issues when all Standards are Included in The Assessment Specifications.   | 69  |
| Figure 3.5: Case 2: The Relationship among Standards, Assessment Specifications and Assessment regarding Validity and Alignment Issues when only a Sample of all Standards are Included in the Assessment Specifications.                                  | 70  |
| Figure 3.6: Case 3: The Relationship among Standards, Assessment Specifications and Assessment Regarding Validity and Alignment Issues when a sample of all Standards are Included in the Assessment Specifications as well as other Knowledge and Skills. | 70  |
| Figure 3.7: Horizontal and Vertical Alignment within an Education System   | 78  |
| Figure 4.1: The Washback Effect of Public Examinations   | 89  |
| Figure 4.2: Basic Model of Washback  | 106 |
| Figure 4.3: Proposed View of Washback: A Curriculum Innovation Model.  | 112 |
| Figure 4.4: Washback Causes and Effects  | 114 |
| Figure 5.1: Hypothesised Washback of the rSECEE on Grade 12 EFL Classrooms   | 121 |
| Figure 5.2: Figure 5.2: Explanatory Sequential Design  | 124 |
| Figure 5.3: Methods and Research Questions   | 139 |
| Figure 6.1: A Sample of a Reviewer’s Coding Responses to rSECEE Assessment Items   | 153 |

|  |     |
|--|-----|
| Figure 6.2: Illustration of How Webb’s Alignment Criteria Measure the Correspondence between Libya’s EFL Curricular Content Standards and rSECEE | 154 |
| Figure 6.3: Distribution of DOK Level of SECEE Test Items by Standards   | 176 |
| Figure 7.1: Phase II Data Collection and Analysis  | 207 |
| Figure 7.2: Grade 12 EFL Classroom Content and Activities  | 231 |
| Figure 7.3: Frequency of Classroom Activities of the Three Grade 12 EFL Classrooms   | 236 |
| Figure 7.4: Grade 12 Students’ BAK towards the rSECEE  | 252 |
| Figure 7.5: Alignment as Links between the Components of an Education System   | 289 |
| Figure 7.6: Alignment as Links between the Components of the Libyan Secondary Level Education System   | 289 |
| Figure 7.7: Alignment as Links Formed by Libyan Teachers Between the Components of the Libyan Secondary Level Education System                   | 294 |
| Figure: 7.8: Revised View of Washback  | 296 |
| Figure 8.1: A Model of Linkage between the High-Stakes Test and Instructional Practices  | 307 |

**\*Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes.**

## List of Abbreviations

- AERA: The American Educational Research Association
- APA: American Psychological Association
- BCE: Before Common Era
- BOR: Balance of Representation
- CAEL: Canadian Academic English Language
- CIV: Construct Irrelevant Variance
- CRM: Cognitive Rigor Matrix
- CUR: Construct Under Representation
- DOK: Depth of Knowledge
- EFL: English as a Foreign Language
- IELTS: International English Language Testing System
- MCQs: Multiple Choice Questions
- NCME: National Council of Measurement in Education
- ROK: Range of Knowledge
- rSECEE: Revised Secondary Education Certificate Examination of English
- SECE: Secondary Education Certificate Examination
- SMEs: Subject Matter Experts.
- TOEFL: Test of English as a foreign language
- TOEIC: The Test of English for International Communication

*“If you wish to converse with me, define your terms”*  
(Voltaire, 1694-1778).

## **Glossary of Terms**

The key constructs informing the research are defined below.

***Alignment:*** Similar to Hansche, this study considers alignment to be “the degree to which expectations (i.e., standards) and assessment are in agreement and serve in conjunction with one another to guide the system toward students learning what they are expected to know and do” (Webb, 1997, p.4).

***Assessment:*** Assessment is a systematic procedure for collecting and synthesizing information regarding students’ achievements (Bertenthal & Wilson, 2005; Popham, 2003). As used in this dissertation, the terms ‘assessment’, ‘examination’ and ‘test’ are synonyms to describe “a formal attempt to determine a student’s status with respect to an educationally relevant variable” (Popham, 2003, p.2), and to measure the target set of content standards. In this dissertation, item is used for the smallest part of an assessment.

***Balance of Representation:*** Is a Webb criterion that judges the extent to which assessment items are evenly spread and distributed across standards (Webb, 1999, 2002).

***Communicative Language Teaching:*** Is a language teaching approach that stresses the importance of promoting learning within authentic contexts, and the emphasis is on the communicative and social aspects of the target language, rather than on its linguistic aspect (Larsen-Freeman, 2000).

***Content Standards:*** Broad descriptions of what students are supposed to both know and be able to do (Popham, 2003).

***Curriculum:*** Within the education literature it is argued that the term *curriculum* can have a range of definitions (Beauchamp, 1982; Jackson, 1996; Kliebard, 1989). In fact, the definitions

of curriculum are almost as numerous as the researchers who write on the subject (Fox, 2004). These definitions range from the simple to the very complex and range from “everything that goes on in schools” (Windh & Gingell, 1999, p.52) to specific documents describing the standards (Donn, 1994). However, most scholars and educators would agree that the curriculum involves a body of content knowledge to be acquired in a particular manner or form (Au, 2007). following Au (2007), in this dissertation, the curriculum is manifested in textbooks, material and internal classroom tests.

***Curricular innovation:*** Green (2007) defines curricular innovation as a “managed process of development whose principal products are teaching (and/or testing) material, methodological skills, and pedagogical values that are perceived as new by potential adopters” (p.46). This definition reflects my own perspective as a programme evaluator.

***Categorical Concurrence:*** Is a category that is concerned with comparing the similarity between the content standards’ expectations and the actual assessment under investigation in alignment research (Martone & Sireci, 2009; Webb, 1999).

***Depth of Knowledge:*** Is the complexity of knowledge that components of an educational system such as standards, assessment and instruction require from students. Webb’s (1997, 1999) DOK level identify four different ways in which one can interact with content. Each level does not just provide the type of thinking used, instead it provides and depends greatly on how “deeply students understand and engage with the content in order to respond” (Hess, 2014, p.1).

***Depth of Knowledge Consistency:*** Is one of Webb’s four categories that have been the focus of analysis in alignment research. It involves the degree of alignment between the cognitive demands of both the standards and the test items (Lane, 2004; Martone & Sireci, 2009; Webb, 1999).

***Hidden Curriculum:*** Unlike the written curriculum, the hidden curriculum is “characterized by its informality and lack of conscious planning”, and includes “values, intergroup relations and collaboration that enables students’ socialization process” (Kentli, 2009, p.83). The hidden curriculum establishes itself by defining: the structure of the instructional activities, social interactions, relationships (Minarechová, 2012).

***High-stakes testing:*** As used in this dissertation, a test can be classified as *high-stakes* when important consequences are linked to students’ performance on the test (Madaus, 1988, Marchant, 2004; Paris, 2000). On the contrary, with low-stakes tests “[t]here is little or no direct formal academic or meaningful consequence for the individual student and other stake-holders” (Abdelfattah, 2010, p. 159). What differentiates a high-stakes test from a low-stakes test is not its form (such as how the test is designed) but its function, i.e., how the test results are used (Cole, Bergin, & Whittaker, 2008). High-stakes testing normally involves national- or state-wide standardised achievement tests (Marchant, 2004). Language proficiency tests such as International English Language Testing System (IELTS), Canadian Academic English Language (CAEL), The Test of English for International Communication (TOEIC), and Test of English as a foreign language (TOEFL) also fall under the high-stakes testing category. In this dissertation, the researcher is concerned with high-stakes national testing, the type of testing that is given to all secondary school students within a country. In accordance with Corbet and Wilson (1991) the definition of high-stakes testing in this dissertation is extended to include tests that “cause the public to make an assessment of quality of their school system which, in turn, can lead to increased public pressure to improve tests scores” (Chapman & Snyder, 2000, p.458).

***Hit:*** In this dissertation, the term ‘hit’ is used to indicate a content standard that has been aligned to an assessment item.

***Innovation:*** Innovation within this study is not considered or interpreted as being constantly innovative and creative in terms of pedagogical approaches or always seeking novel activities to enhance learning. It is instead considered from the perspective of implementing approaches and concepts of teaching that are not part of the Libyan culture of learning, which are nonetheless foreign and new to the teachers' beliefs, assumption and knowledge (BAK, Woods, 1996). For example, the employment of a student-centered approach by teacher participants could be seen as innovation and change because it is not part of the prevailing Libyan culture of learning.

***Mixed-Methods Research:*** Mixed-methods (MM) research is an approach that guides the collection and analysis of both qualitative and quantitative data in a single study (Creswell and Plano Clark, 2007).

***Performance Standards:*** Broad descriptions of the level the students should attain of the set knowledge and skills defined by the content standards (see Standards Below).

***Range of Knowledge:*** Is a consistency category in Webb's model of alignment research that considers on the breadth of standards as compared to the breadth of assessment (Webb, 1999, 2002).

***Secondary Education Certificate Examination of English (SECEE):*** A high-stakes, external examination that is administrated by the Libyan Ministry of Education to Libyan students at the end of Grade 12. Libyan students are required to demonstrate their learning of the subject matter knowledge by taking this test. The results of this test make life changing decisions for Libyan Grade 12 students.

***Standards:*** As used in this dissertation, standards define the "goals for student learning and focus the attention of teachers, students, parents and all the others concerned with education" on what students need to know, understand, and be able to do (Bertenthal & Wilson, 2005, p.2).

Different countries use different labels or terms for identifying levels of educational expectations. In this dissertation, the conventions of content standards and ‘content objectives’ are employed to describe two levels of expectations that students are expected to know and achieve. Content objectives in this dissertation are treated as more detailed descriptors of what students are expected to know and do in order to demonstrate their attainment of the standards; hence, they are a more detailed specification of the standards.

**Teaching:** Teaching is another component of any education system. It is defined in this study as the process of helping and guiding students to acquire the intended knowledge and skills by means of instruction (Safritz, Koppe, & Soper, 1988). This, in turn, implies that teaching should include all kinds of instruction that can be offered to the students to provide them with a reasonable opportunity to learn.

**Thinking skills:** The broadest view of thinking as described by Dewey (1933) includes anything that passes through a person’s mind. However, thinking in its most refined sense as referred by Dewey (1933) is reflective thought. Reflective thought is commonly known as higher order thinking (Hmelo & Ferrari, 1997). Resnick (1987) argues that higher order thinking skills involves complex, non-systematic, distinctive judgements, and the considerations of multiple sources and solutions. In this dissertation, thinking skills are defined as cognitive operations or process that can be building blocks of thinking (Kelly, 2011). We use our thinking skills when we “try to make sense of the experiences, organise information, make connections, ask questions, make plans, or decide on what to do” ( Kelly, 2011, p.1).

**Transmission teaching:** In this model of teaching the teacher just prepares the delivered content and transmits it to learners. The learners’ “role is to receive, store”, and “to commit the facts and

procedures to memory and strive to become fluent and precise” (Tishman, Jay, & Perkins, 1993, p. 149).

**Washback:** In applied linguistics and language testing there are numerous definitions of washback. These definitions range from the simple to the very complex, differing in terms of scope and intentionality (Green, 2013). This study considers washback to be “[t]he influence of testing on teaching and learning” (Bailey, 1996, p.259), and “[t]he extent to which the test influences language teachers and learners to do things that they would not necessarily otherwise do” (Messick, 1996, p.24).

**Validity:** Drawing primarily on Messick’s (1989) reconceptualization, validity in this dissertation is defined as “the degree to which accumulated evidence and theory support a specific interpretation of test scores for a given use of a test” (APA, AERA, NCME, 2014, p. 193).

## Chapter I

### Introduction

*“Education is the occupation of children and adolescents around the world; it is a moral imperative for some, a social and economic opportunity for others, an adventure in curiosity and learning for some, and a duty for most” (Paris, 2000, p.1).*

#### 1.1. Background

Paris (2000) argues that education and the schooling systems around the world share common features regardless of different cultures and contexts. He notes that the daily school routine, teacher roles, the overall school environment, the sources of knowledge (for example, textbooks), and the day-to-day demonstrations of mastery can be found in almost all schools around the world. ‘The test’ and all the related testing practices, as noted by Paris (2000), are the most universal components of all education systems. For more than a century most of the world has been administering tests, often devised and scored externally for social accountability, or to provide information to assist decisions on an individual’s future, reward and sanctions (Cheng, 2008; Ross, 2008; William, 2001).

A test can be described as *high-stakes* when its scores are used to make important decisions that have important consequences not only for the test takers but also for teachers, schools, and society in general (Heubert & Hauser, 1998; Madaus, 1988; Wall, 2012). The context of high-stakes testing often include the test being administered to large groups of test takers within a set time frame; paper-and-pencil or computerised formats, which are answered by shading circles of multiple choice items on a machine-scored/scantron forms or by clicking the correct answers on

a computer screen (Fox & Cheng, 2017). Furthermore, the prevailing assumption among policy-makers and other stake-holders such parents is that “higher scores [on high-stakes tests] indicate a higher quality of teaching and learning” (Paris, 2000, p.2). It is worth noting that most high-stakes tests are norm-referenced. In essence, it measures how well an examinee did on the test in comparison to a large group of test takers (Chapman & Snyder 2000; Marchant, 2004). For example, a student achieving “Excellent”, or “Good” is relative to other students at the same grade level. Currently, high-stakes standardised tests produced by testing companies prevail as the “scientific and objective indicator of students’ knowledge” (Paris, 2000, p.1).

Importantly, high-stakes testing is not a new phenomenon in education (Minarechová, 2012). It has been part of many countries’ education systems including China (Cheng, 2008), England, United States of America, Australia, post-communist countries (Chapman & Snyder (2000; Minarechová, 2012; Paris, 2000; Polesel, Dulfer, & Turnbull, 2012), and many Middle Eastern and Gulf countries. Furthermore, high-stakes tests are part of policies that link test scores to either high-school graduation, or, in other scenarios, to teacher salaries, school or state funding and ranking (Orfield & Wald, 2000; Schneider & Ingram, 1997). In other words, the stakes for both teachers and students can include financial awards, recognition, and negative sanctions. Policy-makers do in fact acknowledge the power of tests and use them as “effective tools for controlling educational systems and prescribing the behaviour of those who are affected by their results, administrators, teachers and students” (Shohamy, Donitsa-Schmidt & Freeman, 1996, p.299; see also Chapman & Snyder, 2000; Cheng, Sun, & Ma, 2015; Shohamy, 2007). In fact, ‘it would not be too much of an exaggeration to say that evaluation and testing have become the engine for implementing educational policy’ (Petrie, 1987, p. 175).

Tests and testing practices may be viewed as the power tools of education policies (Shohamy, 2007); they operationally define the beliefs, assumptions and knowledge (BAK, 1996) that are articulated in a curriculum. Within the education literature the term *curriculum* can have a range of definitions (Beauchamp, 1982; Jackson, 1996; Kliebard, 1989). However, most scholars and educators would agree that the curriculum involves a body of content knowledge to be acquired in a particular manner or form (Au, 2007). This definition only considers the level of content, and according to Au (2007) to stop at this level of conceptualization obscures the additional aspects of curriculum, as subject matter content in educational programmes necessitates both the selection and transmission of knowledge. Similarly, Apple (1995) states that a definition that considers all aspects of the curriculum is one that accounts for: the structure of knowledge embedded in a curricular form; the way in which knowledge is organised and presented within a curriculum; and the intended way(s) in which the selected content is delivered (i.e., pedagogy). Accordingly, the three defining aspects of curriculum are subject matter/content knowledge, the structure of curricular knowledge, and pedagogy. In accordance with Au (2007), this fundamental conception of curriculum is the one that the present study employs.

In the context of curricular reform where high-stakes testing plays a fundamental role and the decisions related to test performance carry important consequences either in terms of reward or sanctions, the extent of confidence in, and the justifiability of, test score interpretations ought to be comparable (Hargreaves, Earl, & Schmidt, 2002; La Marca, 2001). In other words, the greater the consequences of a large-scale tests, the greater the need for evidence of test validity (La Marca, 2001, Shohamy et al., 1996; Zumbo, 2009).

In the educational measurement literature, test influence is variously known as ‘test impact’ (Bachman & Palmer, 1996; Baker, 1991; Wall, 1997), ‘consequential validity’ (Messick, 1989,

1996), and measurement driven instruction. Measurement driven instruction refers to “the notion that tests should drive learning” (Shohamy, 1993, p.4). Test impact is a term that refers to the effects that tests have on individuals (teachers and students), on educational systems, and on society at large (Bachman & Palmer, 1996; Wall, 1997). Similar to general education, research in the language testing field has presented evidence that language examinations may have an effect on both teaching and learning; a phenomenon commonly known as *washback* (see, for example, Alderson & Wall, 1993; Cheng, 1997, 1998, 2005; Green 2007). In applied linguistics and language testing there are numerous definitions of washback. These definitions range from the simple to the very complex and differ in terms of scope (Green, 2013, 2014). Given that the focus of this dissertation is the classroom, the following definitions of washback capture the meaning that is employed herein, namely, “[t]he influence of testing on teaching and learning” (Bailey, 1996, p.259), and “[t]he extent to which the test influences language teachers and learners to do things that they would not necessarily otherwise do” (Messick, 1996, p.243). within the field of language testing, Shohamy (1993) focuses on the “connection between testing and the teaching syllabus” (p.4), i.e. curricular alignment.

In an ideal world and to address issues related to the effects of testing, educators and researchers ought to be able to determine whether or not what is assessed in mandated tests aligns with the content standards and classroom instruction (Martone & Sireci, 2009). Alignment can be generally defined as the match between two or more things. Webster’s New World College Dictionary defines the verb align as bringing “into a straight line; to bring into agreement, close cooperation”. In this context, “[i]n an aligned [education] system of standards and assessment, all components are coordinated so that the system works toward a single goal: educating students to reach high academic standards” (Hansche, 1998, p.21). Similar to Hansche,

the current study considers alignment to be “the degree to which expectations (i.e., standards) and assessment are in agreement and serve in conjunction with one another to guide the system toward students learning what they are expected to know and do” (Webb, 1997, p.4). As for assessment, it is defined as a systematic procedure for collecting and synthesizing information regarding students’ achievements (Bertenthal & Wilson, 2005; Popham, 2003). As used in the current study, the terms ‘assessment’, ‘examination’ and ‘test’ are synonyms to describe “a formal attempt to determine a student’s status with respect to an educationally relevant variable” (Popham, 2003, p.2), and to measure the target set of content standards. Assessment is discussed fully in Chapter Three (see Section 3.1.2).

Researchers emphasize that alignment research is one possible means for demonstrating or assessing the link between the three components of any education system: standards, teaching, and testing (Biggs, 2003; La Marca et al., 2000; Martone & Sireci, 2009; Webb, 1999). If these three components work in harmony to deliver a consistent message about what is valued in the education process (Webb, 1999), the students may then have a better opportunity to learn and truly demonstrate what they have attained (Anderson, 2002; Biggs, 2003; Farenga, Joyce, & Ness, 2002; Martone & Sireci, 2009). However, the lack of alignment between components of an education system may result in *negative washback* (Green, 2007, 2013, 2014; Linn, 2000; Shohamy, 1997; Tan, 2008; Wall, 2005). Madaus (1988) asserts that “negative washback definitely results in cramming, narrowing the curriculum”, and focussing attention of those skills that are the most relevant to testing (p.22). This, in turn, means that the students may not develop the whole range of concepts and skills articulated in the curriculum and concomitant content and performance standards, and negative washback may undermine the teaching and learning

experience at the classroom level and substantially affect students' academic performance (Frederiksen, 1984; Popham, 1987).

## 1.2. Purpose of the Study

The large volume of empirical work documenting the impact of high-stakes testing reveals worrying concerns (see for example Fox & Cheng, 2007, 2017; Shohamy, 2007; Paris, 2000). These concerns are particularly evident within the context of the current Libyan secondary education system. My experience as a high school teacher, a full acquaintanceship with the context of the present study, findings from previous research conducted in Libya (Abdulhamid, 2011) and conversations with Grade 12 teachers and students have raised my concern regarding the current Libyan secondary education system and the operation of its high-stakes testing programme.

In addition, the lack of research regarding the Libyan approach contained within the revised Secondary Education Certificate Examination (SECE) suggests that empirical focus is essential. To begin to address this gap in the research literature, this dissertation examines the relationship between the degree of alignment between components of an educational system (i.e., standards and high-stakes testing) and the washback at the classroom level. In order to describe the relationship between the degree of alignment and washback at the classroom level, this dissertation: (1) looks at the degree of alignment between the revised Secondary Education Certificate Examination of English (rSECEE) and Libya's EFL content standards; (2) investigates how the *washback* of the rSECEE operates at the classroom level, that is, how a change in this high-stakes national examination may influence teachers and teaching, and learners and learning. In this dissertation, the focus is on the washback of large-scale, high-stakes test in this foreign language educational context.

At the end of Grade 12, Libyan students are required to demonstrate their learning of different subject matter knowledge by taking high-stakes, external examinations that are administered by the Libyan Ministry of Education. These high-stakes tests make life changing decisions for Libyan students. Test results are used to assign students to different university departments or vocational colleges; determine whether or not students will be promoted to the next grade level; and decide whether students will receive high school certification. It should be noted that the SECEs are norm-referenced tests which rank and compare each Libyan student's performance against other students' performance.

The Libyan education system has been undergoing reform since 2009. The revised English as a Foreign Language (EFL) curriculum and rSECEE aim to produce a positive washback effect on school teaching and learning as they encourage students to be active in the classroom (personal communication with the Curriculum Department, the Libyan Ministry of Education, 2016). This approach to teaching and learning differs considerably from the traditional Libyan teacher-centered and *transfer of knowledge*<sup>1</sup> model that was in place for centuries (see Section 2.4.2 for a detailed description of the model). Similar to other education systems, because of the high-stakes involved, teachers are knowledgeable about the criteria required by these large-scale, external assessments (Cheng & Curtis, 2004; Green, 2007, 2014). In other words, because of the high stakes involved, the SECEE may be exerting considerable influence on both teaching and learning (Madaus, 1988). The rSECEE, like other language tests implemented within education contexts, has been “unquestioned, unchallenged, unmonitored and uncontrolled” (Shohamy,

---

<sup>1</sup> In the “transmission model of learning” student are passive learners, they receive information by teacher lecturing and from the target text. It is argued by Bransford and Schwartz (1999) that students in this model of learning fail to apply what they have learnt in real-life situations.

2007, p.524). With important decisions resting on the results of the rSECEE, it is important to understand how well the test is performing within the Libyan context, and to judge if decisions based on the rSECEE are reflecting accurate interpretations and resulting in best practice.

In accordance with Onaiba (2014), this study is important because within the Libyan context there is a lack of empirical research focussing on issues related to language testing and the influence of examinations on both teaching and learning. In addition, an alignment review of assessment and standards is fundamental whenever assessment is modified, or the passing scores are changed (La Marca, 2001), as is the case within the current Libyan assessment context. Furthermore, from the time the Libyan education system underwent reform, very little or no information is known about the degree of alignment between the rSECEE and Libya's EFL content standards, or how the rSECEE is operating in the classroom and its possible influence on both teaching and learning. Therefore, the empirical evidence derived from this study aims to provide useful information for policy-makers that can be used (if needed) to change assessment procedures, or alter standards and to validate the degree to which these two educational components are directed towards mutual expectations for students' learning (Martone & Sireci, 2009; Roach et al., 2005). In addition, this research aims to inform Libyan policy-makers about the success of their curricular implementations; and enhance and deepen understanding of the washback phenomena (Watanabe, 2004).

The levels of concern raised in the education literature calls for the study to be rigorous and comprehensive involving multiple views, including those of students, teachers, and policy-makers (Polesel et al., 2012). More precisely, the present study focuses on several significant issues including:

- The degree of alignment between Libyan EFL curriculum and articulated content standards, and the rSECEE.
- The impact of the test on teachers' classroom practices and their accounts.
- The impact of the test on students' accounts of their learning.

### 1.3. Research Questions

The present study was guided by the following research questions:

**What is the relationship between the degree of alignment and the washback of the rSECEE? What are the implications of this relationship for key stake-holders (e.g. policy- makers, test developers, teachers, and students)?**

These questions guided the researcher to explore:

- 1) **To what degree is the rSECEE aligned with Libya's EFL content standards?**
- 2) **What is the nature and scope (Cheng, 2004) of the washback (if any) of the rSECEE on the Libyan EFL Grade 12 teaching and learning?**

These four overarching research questions were addressed through the following sub-questions:

- a) What evidence is there of washback of the rSECEE on teachers i.e., how does the rSECEE influence teachers' accounts of teaching and testing (i.e., external and internal testing)?
- b) To what extent does the rSECEE appear to influence teachers' teaching practices?
- c) How does the rSECEE influence learners' accounts of learning?

### 1.4. Overview of the Thesis

Chapter Two provides a brief description of the historical background of Libya and its official languages. An overview of the Libyan education system and the way it functions then follows. In addition, a detailed description of the English language teaching situation in Libya is

provided, along with the development of Libya's secondary schooling system over the past five decades. Finally, the characteristics of the educational and classroom culture, the role of examinations in Libya, and the Libyan SECEE are presented in detail. Chapters Three and Four cover the target research literature. In Chapter Three the definition and functions of three fundamental components of an education system are discussed along with the triangular relationship between standards, testing, and teaching and learning. Finally, this chapter presents an overview of alignment, in which a more detailed definition of alignment and its contemporary research approaches is offered. Chapter Four synthesizes literature from the language testing and education research fields to describe and discuss washback and criticisms of high-stakes testing.

Chapter Five provides an overarching review of the methodology. A more detailed description of the data collection and analysis procedures for Phases I and II are provided in Chapters Six and Seven respectively. The methodology discussed in Chapter Five includes the research questions, presuppositions, researcher's research orientation, an overall description of the research design, the methods used in both phases of the study, and reference to how the reliability and trustworthiness of the research was maintained. Chapter Six presents the specifics of the methods used in Phase I (including a description of the participants, data collection and data analysis procedures). This is then followed by results, and their discussion. Similar to Chapter Six, Chapter Seven presents the specifics of the methods used in Phases II. This is then followed by results, and their discussion. The final part of Chapter Seven synthesizes and merges the results of the two phases into one overall interpretation of the phenomenon. Finally, Chapter Eight concludes the whole study with some implications for test developers, administrators, and policy makers, along with a short discussion on the limitations of the research and suggestions for future research.

## **CHAPTER II**

### **Background**

In order to address the relationship between alignment and washback in the Libyan Grade 12 EFL context, a detailed description of both the macro (i.e., alignment between EFL standards, curriculum standardised textbook, and test) and micro (i.e., washback on teaching and learning at the classroom level) is a fundamental requirement. In order to understand the macro context of this study, in the sections which follow below, a brief description of the historical background of Libya and its official languages is provided first. An overview of the Libyan education system and the way it functions then follows. In addition, a detailed description of the English language teaching situation in Libya is provided, along with the development of Libya's secondary schooling system over the past five decades. Finally, in order to understand the micro context of this study, the characteristics of the educational and classroom culture, the role of examinations in Libya, and the rSECEE are presented in detail.

#### **2.1. Historic Background**

Libya covers an area of 1.7 million sq. km and is located on the North African coast of the Mediterranean Sea. This vast territory is as large as France, Spain, Italy and Germany combined, but 90% is desert (Dughri, 1980). In 2017, it had a population of 6,408,178 (World Population Review, 2017) with 1.126 million estimated to be living in the capital Tripoli. From about 700 B.C.E., Libya was subject to waves of invasions. In 1551, Libya became part of the growing Ottoman Empire. Turkish control, which prevailed until the 1911 Italian invasion, had a long-term impeding effect on the country, in terms of social, economic, educational and political development (Dughri, 1980).

The Italians began their occupation in 1911, but were met with violent resistance from the Libyan population. The Italian government left Libya with a legacy of buildings, roads, ports and other public equipment and mechanisms (Dughri, 1980). However, Libya paid dearly for the occupation. Libyans were forced to leave their farms, deprived of their political rights, left without economic benefits, and not considered in terms of any social reforms. In addition, the Italian colonisation did not promote education, in contrast to the French in Algeria. No Libyan received education or training; consequently, there were shortages of teachers, skilled workers, technicians and administrators (Country Studies, 1987).

In 1943, Libya was occupied by the Allied forces. Tripolitania<sup>2</sup> and Cyrenaica<sup>3</sup> were governed by the British and Fezzan<sup>4</sup> by the French. The Italians were defeated in 1945 at Tobruk<sup>5</sup> and Libya's relationship with Britain became stronger. From 1945 to 1951 Libya was administrated by Britain. As a result, Libyan-British affairs strengthened, and new business, trade and industry were generated between the two nations. As a result, English increasingly became the language of business (Blackwell, 2003). This Libyan-British relationship became even stronger after Libya gained full independence in 1951.

In 1955, oil exploration started and oil was first exported in 1961. The discovery of oil transformed Libya from being one of the poorest countries in the world to one of the most prosperous in per capita terms. However, the oil discovery did not solve all the problems that Libya encountered during the 1960s. As Libya grew wealthier there was a growing resentment within the nation itself. On the 1<sup>st</sup> of September 1969, a revolution took place led by Colonel

---

<sup>2</sup> The historic region of western Libya centered on the coastal city of Tripoli.

<sup>3</sup> The historic region of northeastern part of modern Libya.

<sup>4</sup> The Fezzan region touches the Libyan coastal zone in the north, the Libyan Desert in the east, the Tibesti massif of Chad to the south, and the Hoggar massif along with the Grand Erg Oriental of Algeria to the west.

<sup>5</sup> A city located on the eastern Mediterranean coast of Libya and on the borders of Egypt.

Muammar Al-Gaddafi (henceforth Gaddafi). His regime lasted 42 years until 2011. The 1969 revolution (known as the Al-Fatah Revolution) resulted in changes to Libyan society, through the implementation of major political, social and economic reforms. “Political reform was at the heart of all changes” (Maghur, 2010, p.1).

The Al-Fatah<sup>6</sup> revolution was based on three underlying ideologies: socialism, freedom and unity. Then, in the early 1970s, Libya announced a cultural revolution. “Other countries, besides Arab-Islamic ones, were considered as imported cultures which were to be rejected, even opposed” (Maghur, 2010, p.5). By 1970, the Libyan government declared that both the American and British army bases were to be closed. Gradually by 1971, all libraries and any cultural association administrated by any foreign government were also closed. Foreign books, magazines and music were prohibited, and private foreign schools were also closed.

Between 1984 and 1999 Libyan-British and Libyan-American relations deteriorated. Three major incidents affected the Libyan international political status. These were: (1) the killing of the British police woman Yvonne Fletcher<sup>7</sup> in April 1984; (2) a bombing in West Berlin in 1986,<sup>8</sup> and (3) the Lockerbie<sup>9</sup> incident in 1988. As a result, the US and UN imposed sanctions on Libya in 1986 and 1992 respectively. Due to the US sanctions (1986-2004) and the UN sanctions from 1992 to 1999, Libya was isolated, which affected the country dramatically. The resulting sanctions and trade embargoes caused import costs to increase, stoking inflation, and subjecting Libyans to a declining standard of living (Country Profile, 2005).

---

<sup>6</sup> The noun Al-Fatah has not been modified by a definite article ‘*the*’ because in the Arabic language this noun has already been pre-modified by ‘*Al*’.

<sup>7</sup> She was killed whilst on duty during a protest by Libyan opposition outside the Libyan embassy in London, allegedly by a shot from inside the embassy.

<sup>8</sup> A bombing at a nightclub visited by the American military employees allegedly carried out by Libyan operatives, to which the US responded with an air raid on Tripoli in 1986.

<sup>9</sup> Pan Am Flight 103 was bombed over Scotland, which some blamed on Libyan operatives.

It was not until 1999 that the UN suspended its sanctions regime after Libya handed over the two suspects of the Lockerbie bombing and in 2003, its eleven-year sanction regime was completely lifted. Furthermore, in 2004, the US lifted many of its sanctions. Accordingly, Libya re-emerged in the international political context, and established new political, commercial, and business links with other parts of the world (Country Profile, 2005). In 2011, a revolution broke out against the brutal Gaddafi regime, and since then the official name for the ‘Great Socialist People’s Libyan Arab Jamahiriya’ has reverted to Libya.

## **2.2. Languages**

Arabic is the main language spoken in Libya, although different forms of Arabic used: Classical Arabic, Modern Standardised Arabic (MSA), and the Libyan Arabic Dialect (Bagigni, 2016). Classical Arabic is the language of the Quran and daily Islamic rituals. MSA is a much-simplified form of Classical Arabic. It is a prestigious standard that is used in government affairs, the press, media, and education. However, the Libyan Arabic Dialect is the informal Arabic used on a daily basis. Both accent and the dialect differ from one city to another. The mother tongue dialect that children first hear is the Libyan Arabic Dialect; however, as soon as a child enters school he/she will be exposed to MSA as this is the medium of teaching and the language of textbooks (Bagigni, 2016).

It is worth noting that Arabic is not the only spoken or written language used in Libya, Berber (with its different varieties) is also used. The Berber minority in Libya live in Nafusa Mountain, Zuwara, and some towns in Ghadames and Sowknah (Bagigni, 2016). Only five percent (approximately 150,000 in 2016) of the Libyan population speak this language (Bagigni, 2016), and Berber does not have any official status within Libya.

## 2.3. The Evolution of the Libyan Education

### 2.3.1. Development of Education before Independence

Unfortunately, the history of education in Libya is not adequately documented and little information on the development of Libya's educational institutions exists. This short history is based on the limited documentation available.

Education during the Islamic period followed the Muslim education system as in Baghdad, Cairo and Tunisia. The *Quran* (the central religious text of Islam) was the foundation stone of Islamic education. In Libya, the Islamic schools came to be known as *Kuttab*s.<sup>10</sup> The *Kuttab* was only open to males, up to the age of fifteen. After completing school, they would leave to take on practical work experience and training on their fathers' farms or engage in other technical work (Dughri, 1980).

In 1939, during the Italian colonisation, there were only 6,736 students enrolled in 64 primary schools across the whole country. Libyans were not entitled to education beyond the primary stage, because they were considered inherently inferior to Italians. However, education started to improve somewhat during the Allied occupation. In October 1943, the British government opened new schools. The Libyan education system at that time faced a lack of qualified teachers; there were only 170 teachers in Libya in 1943 according to Dughri (1980), and thus any large scale of improvement during those years was hindered. At the time, the Palestinian curriculum was used in Tripolitania and Egyptian in Cyrenaica. However, in 1949, Tripolitania adopted the Egyptian curriculum as well. It was not until 1951 after Libya's

---

<sup>10</sup> In *Kuttab*, the location would be a mosque, the student would memorise passages from the Quran and write them on his wooden board. After the passage has been memorised the student would erase it and learn the subsequent passage. This would be an ongoing process until the whole Quran was learnt by heart.

independence that a Libyan curriculum was implemented across the education system for the first time (Dughri, 1980).

### **2.3.2. Development of Education since Independence**

At the time of independence, fewer than 10% of the adult population had received education in Libya. Thus, the newly independent government began by focussing on the fundamentals for an appropriate school system and “opened the doors to education for a large number of students in the shortest possible time” (Dughri, 1980, p.35). In 1951, the Libyan government formally recognised that education should be the right for every Libyan, and education became both compulsory (up until secondary level) and free. Education at that time was seen as a critical factor for contributing to both the social and economic development of the country. In 1954, the official census confirmed that 81.1% of the Libyan population was illiterate (Otman & Karlberg, 2007). However, by 1969, 50% of children from six to eleven years of age were enrolled in primary schools. The 1969 revolution, and the establishment of modern Libya, brought with it changes (albeit limited) to the education system. Primary schools increased in number and the importance of education was further emphasised. By 2004, the figure for illiteracy had fallen dramatically to 12.6% (Otman & Karlberg, 2007).

Education became the right for every Libyan citizen, and is compulsory for both genders from age six until the secondary level (ninth grade). The Libyan education system is highly centralised (Elabbar, 2011) and policy is strongly influenced by the governing political system (Aloreibi & Carey, 2017). The General People’s Committee for Education, (since the 2011 revolution this is known as the Ministry of Education) is the highest administrative power in Libya with regards to education. Abou Jaafar (2003) asserts that the responsibility of the Committee was to “set education and training plans, provide teachers, trainers ... to meet the

training needs of different sectors and organize and execute programmes in schools” (p.16). The Secretary of the Committee (since 2011, renamed the Minister for Education) is the person in charge of the education system. This implies that he (with the help of the committee members) is in charge of developing the nation’s education policy (Mohamed, 1987). In addition, the planning of the syllabus and the choice of textbook resources are typically accomplished by the expert inspectorate. Each inspectorate chooses the syllabus and textbooks for their assigned subject. The professional inspector usually has control over the decision-making process. Additionally, the syllabus in Libya is characterised by its stability. When something is approved, it is expected to continue to function for an extended period. The system is homogeneous across Libya; similar arrangements and similar requirements in terms of curriculum and textbooks are implemented in all schools. Not only are the same textbooks used, but also the same pedagogy is adopted throughout the country (Aloreibi & Carey, 2017; Mohamed, 1987). Importantly, to the best of my knowledge, the Libyan education system still follows the same homogenous schooling process.

Education in Libya today is delivered in two ways, public and private. Public education is free and is funded by the Ministry of Education. Within the private sector, schools are run by their principals and administrative staff and guardians pay for their children to be educated. The majority of Libyans use the public education system because it is free of charge, and many families have more than one child attending schools (El Abbar, 2016). Public-school education in Libya is currently organised on a 12-year basis.

In sum, the Libyan education system can be described as a centralised system that follows a top-down policy under the supervision of the Ministry of Education. The system provides 12 years of compulsory education. Six years are primary schooling (from age 6 to 11), three years is

lower-secondary schooling (previously known as preparatory, from age 12 to 15), and three years secondary schooling (from age 15 to 18). Those students who successfully complete the secondary schooling stage have the option of completing tertiary education (4 years). University programmes follow a semester path, comprising eight to twelve semesters. Those who are unsuccessful at completing the secondary school level have the option of entering the job market or attending career-oriented schools, such as technical and vocational schools that provide training in areas like computer sciences, engineering, construction, or mechanics. Students attend schools five days per week (from Sunday to Thursday) and have six lessons per day (45 minutes per lesson). Figure 2.1 depicts the possible routes that students can follow during their schooling in Libya.

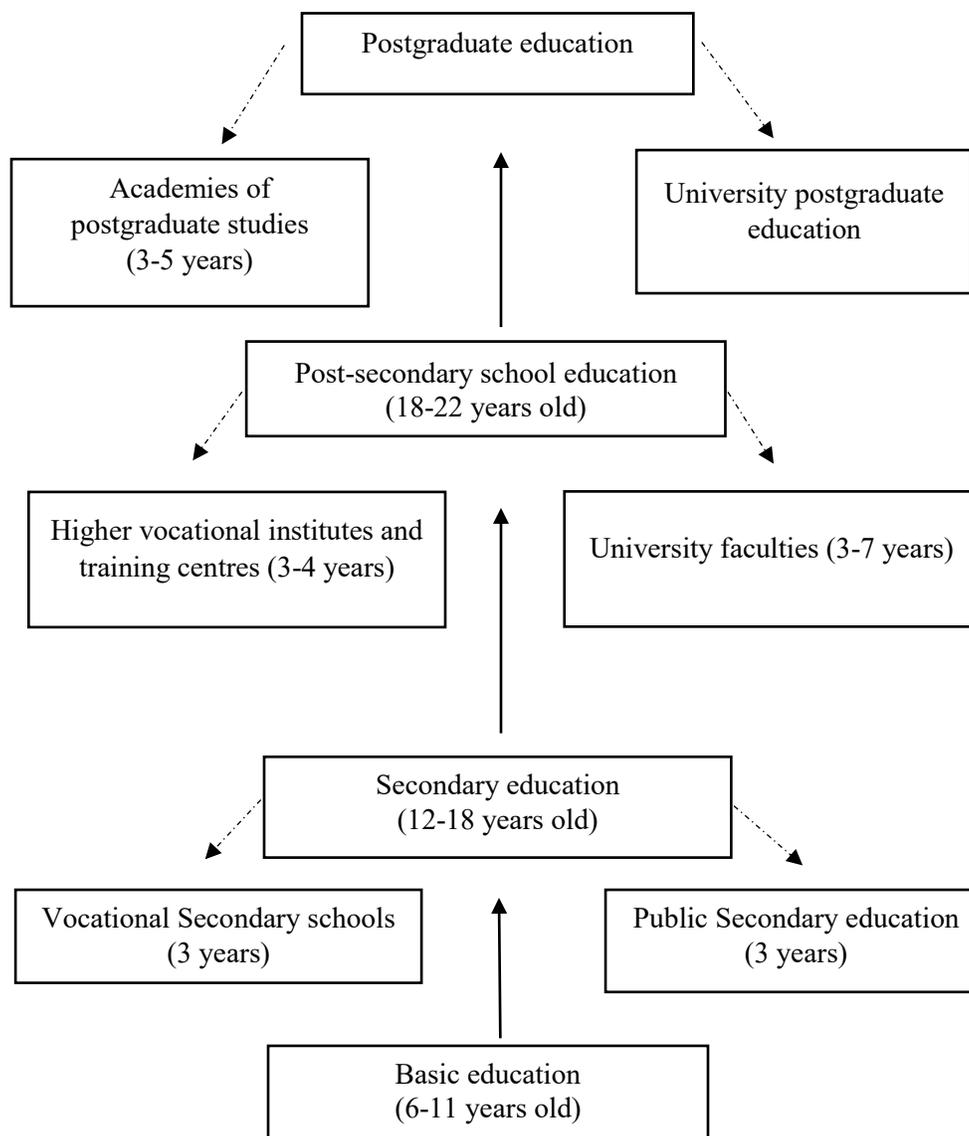


Figure 2.1: Structure of the Formal Education System in Libya  
Source: Adapted from Asker, (2012)

#### 2.4. English Language Teaching in Libya

The teaching of English in Libya dates back to the mid-1940s. As soon as the British began to govern the northern part of the country (Tripolitania and Cyrenaica), intensive English language programmes were organised, for whoever was interested in learning and acquiring the language. The initial English language series implemented in Libya was *Basic Way to English* by

K.C. Odgen (Hashim, 1997), and the Basic Reading Books 1 and 2 by L.W. Lockart (Magoub, 1977). The books were based on the philosophy of using frequent vocabulary and facilitating the language learners' control of language structures within the scope of 850 words. The dominating teaching approach of that time was grammar-based<sup>11</sup>. Memorisation of facts, vocabulary and grammatical structures were the focus of their L2 instruction. Reading and writing were also emphasised as they were regarded as the basic skills of language (Hashim, 1997). In these classes, language development was pictured as the formation of habits and memorisation of structures. The focus was on the form (or the ability to produce a grammatical response), rather than the meaning or general communicative ability of the student. This English language teaching (ELT) pedagogy aligned with the traditional educational cultures in Libya that are discussed in the following section. Assessment was administrated periodically, and was based entirely on the assigned textbook content (Magoub, 1977).

From 1954, after the signing of the Anglo-Libyan treaty, English continued to be taught from Grade 5 (the year before the end of primary level education), until the end of lower-secondary education (Grade 9). In the 1960s, Libyan education officials along with the English language teachers and inspectors stressed the need for a new English language programme. They felt a need for a programme that would change the whole course of English language pedagogy in Libya. Consequently, this led the introduction and implementation of a new series known as *English for Libya*, developed by an inspector called Mustafa Gusbi. This series, which sought to meet the Libyan language learners' linguistic and social needs (Hashim, 1997), was well-established in the Libyan curriculum at the lower-secondary level (Grades 7, 8 and 9). Mahgoub

---

<sup>11</sup> Grammar-translation method of language teaching is an approach where the teacher presents language structures, then practiced in the form of exercises by the learners (Chang, 2011).

(1977) contends that the Gusbi series represented “the first serious attempt to bring up to-date methods and materials into the classroom of the preparatory schools of [Libya] and to improve considerably the language taught there” (p.4).

Hashim (1997) cites that the Ministry of Education target objectives of teaching English in the lower-secondary level during the Gusbi series as follows:

The aim of teaching English in the preparatory level is to enable the learners, in three scholastic years, to understand spoken English and to speak the language fluently as well as to read and comprehend English texts of basic vocabulary and common structures and to be able to write a number of sentences on a certain subject. In this stage focus should be directed to all language skills since these are considered indivisible and integrative. A learner who cannot continue, his or her, education beyond this stage for one or another reason would find what he or she had learned of the English language could be helpful in his or her professional career. For those who would continue and went to secondary schools, would find that they had a solid background in the English language that qualify them for advanced levels in learning the language.

The Gusbi series was based on the audio-lingual methods of Lado and was fully contextualized to the Libyan context. The Lado approach was based on the belief that language learning is habit formation (Fox, 2011, personal communication). Drilling, repetition and over-learning were core elements of the employed teaching approach. The names employed within the Gusbi series were Libyan, pictures and customs were also Libyan, and the emphasis was on spoken English (“Listen” and “Repeat”). Patterns practiced through oral-aural drills and scripted role play dialogues were the same as those in the textbooks (Mahgoub, 1977). In order to implement this program at school level, thousands of native-speaking teachers (largely from the

Peace Corps in the US) were brought to Libya to teach. There was a similar initiative for French during this period and many French speakers met the requirement for service in the military by teaching French in Libyan classrooms (Fox, 2018, Personal Communication).

Hashim (1997) stated that in the 1970-71 academic year, the series *English for Libya* was revised by Gusbi in conjunction with R. John and was titled *Further English for Libya*. This series was intended to be an extension of the previous series *English for Libya*. It incorporated two textbooks, titled *Book 1* and *Book 2*, and was planned for the first and second year of secondary school education. However, in third year different textbooks were employed to match the student's literary and scientific specialisation. It is worth noting that the Libyan Ministry of Education considered English competence to be vitally important for students wishing to study at the tertiary level. In essence, this reflects the spread of English as a lingua franca of science and technology (Bagigni, 2016).

However, after the Cultural Revolution in 1973, English was not taught after Grade 7 (the first year of lower-secondary level education). Within the Cultural Revolution Act, as stated earlier, the Gaddafi regime closed down all cultural associations administered by foreign governments and foreign language schools. Consequently, the Revolutionary Council revoked the contracts of all the foreign language teachers (English and French) who were forced to leave the country. However, until the mid-1980s English remained part of the school curriculum in Libya but there was a change of policy in 1986, which reduced the predominance of foreign teachers in Libya at that time. This was accompanied by an Arabisation campaign (in response to the American air raid, and the US sanctions on Libya). Consequently, in 1986 ELT was withdrawn altogether from the Libyan education system; this lasted until 1992 (Elmabruk, 2008). Subsequently, the government banned the teaching of all foreign languages in schools,

specifically English and French. School curricula were reorganised in favour of a renewed emphasis on scientific subjects, humanities, Arabic language and Quranic education, at the expense of English.

In 1986, regular students in Grade 7 had had no English language instruction. However, those who were lucky to have developed some English despite the policy could not improve beyond their level if they were in secondary education. Although English was banned in basic and lower-secondary education, it continued to be taught and was the medium of instruction in many tertiary-level institutions, such as for medicine, pharmacy, and engineering. Consequently, the gap in teaching English created enormous problems. Students who had enrolled in language faculties had limited linguistic competence, and in extreme cases no linguistic competence at all, and, as a result, students in scientific and technical faculties found their studies extremely challenging (Maghur, 2010). The main long-term effect of the banning of English in Libya resulted in a whole generation of university graduates having little or no knowledge of English (Elmabruk, 2008; Maghur, 2010). Elmabruk (2008) further argues that the “consequences of this ill-advised withdrawal of English were far reaching, not just for learners, but also for teachers and inspectors alike” (p.23). From 1986-1992, ELT teachers had to take on different jobs, which included teaching geography, history, and Arabic. In 2007, Tripoli University<sup>12</sup> reported that 80% of English teachers who had left the English teaching profession did not return to the profession after the six years suspension of English was over (Asker, 2012).

The effects were even more striking when the Libyan government re-emerged as an international actor in 1997, and began to promote foreign trade and investment. Many Libyans found themselves unable to integrate into the market due to their lack of English. Otman and

---

<sup>12</sup> Previously known as Al-Fatah University when the Gaddafi regime was in place.

Karlberg (2007) contend that the banning of English “[f]or Libya...proved to be a fundamental and disastrous mistake... [it] has set back Libya, in terms of educational quality, by two generations” (p.110).

When diplomatic relations with the West began to improve in 1999, English resumed its position of importance in Libya as an international language. In addition, other social and economic factors made the re-emergence of English of importance in Libyan society. These included:

1. English is the main medium of communication within the hydrocarbon sector.
2. A flood of foreign businesses establishing regional offices in Libya creating a need for an English-speaking staff.
3. Growth in the tourism sector resulting in the need for local representatives and guides able to communicate with holiday-makers and tour employees.
4. In job advertisements,<sup>13</sup> having English competency has become a fundamental requirement for employment in both government and foreign administrations.

The economic boom in Libya has to a great extent fuelled the demand for English. Hence, English was also booming. However, the Libyan workforce still lacked the necessary language skills, which dramatically affected employment opportunities. For a number of years there have been intensive English programmes taking place, such as sending Libyans for overseas courses (especially the UK), but the policy has not been sufficient to fill the gap in the Libyan market. Thus, Maghur (2010) notes that it is “important to reiterate that in today’s Libyan market it is English and English alone that is a passport for work” (p.6).

---

<sup>13</sup> Libya online job advertisement at: <http://www.libyaonline.com/business/pages.php?cid=300>

In reaction to the educational and market needs, the Libyan government reconsidered its ELT planning policy in 1999. English was reintroduced to Grade Five in the academic year 2006/07. The curricular objectives for students suggest the requirements for the new pedagogical approach, as outlined by Orafi (2008):

- To leave school with a much better access to the world through the *lingua franca* that English has become;
- To create an interest in English as a communication tool, and to help students develop the skills to start using this tool effectively.
- To help students use the basic spoken and written forms of the English language.
- To help students learn a series of complex skills: these include reading and listening skills that help attain meaning efficiently; for example, skimming and scanning and interpreting the message of the text. They also include the speaking and writing skills that help the students organize and communicate meaning effectively.

These objectives guided the introduction of a new textbook series, entitled *English for Libya*. Under the supervision of the Libyan Ministry of Education the series was published by Granet Ltd, Reading, U.K. The material was designed by a team of English native speakers. For the basic schooling and lower-secondary level, the programme consisted of five levels, from grades five to nine. The series for the secondary level consisted of three levels from grade 10 to 12. The secondary level series of the *English for Libya* was tailored as “a multi-level formal subject specific EFL courses intended to be taught to students” within specialized streams (Bagigni, 2016, p.47). Many Libyan researchers including Ali (2008), Gumah (2011), Omar (2014), Orafi (2008), Orafi and Borg (2009), and Tantani (2012), document that the revised EFL material (i.e., textbooks) for the years 2004 to 2013 reflect a curriculum that was *communicatively oriented*. In

other words, it encouraged and employed communicative language teaching approaches and pedagogy. The communicative approach stresses the importance of promoting learning within authentic contexts, and the emphasis is on the communicative and social aspects of the target language (Larsen-Freeman, 2000). The communicative approach generally attempts to place language teaching within real-world situations (Larsen-Freeman, 2000). Different from the traditional grammar teaching method, grammar learning in the communicative approach, is “emphasized by communication through the approaches of ‘learning by doing’, through students’ participation or co-operative completion of teaching tasks between or among students and teachers” (Chang, 2011, p. 11).

Contemporary theories of learning (such as Vygotsky) were considered in the presentation of the curriculum (Ali, 2008). Furthermore, the revised ELT teaching material was based on “interaction and communication approaches to encourage learners to learn language through real situations and avoid the traditional methods in teaching a foreign language” (Ali, 2008, p.9). The textbook activities varied from writing formal letters to role play of a story or an event (Orafi & Borg, 2009). Overall, Tantani (2012) argues that the language activities within the revised EFL textbooks aimed at developing learners’ linguistic knowledge (such as grammar and vocabulary) in order to develop the four language skills in the long run (reading, writing, listening and speaking). Ali (2008) further notes that students’ active participation in language classrooms and their concomitant cognitive development was the fundamental focus of the newly revised ELT curriculum.

In addition, EFL textbooks used during this period required Libyan students to communicate in English (orally and in writing) within the classroom context (Omar, 2014). Omar (2014) further states that the teaching approaches and models of learning employed in the revised EFL

textbooks taught in the Specialized Secondary Programme employed problem-solving, critical thinking, and cooperative learning skills. Thinking in its most refined sense as referred by Dewey (1933) is *reflective thought*, which is known nowadays as higher order thinking (Hmelo & Ferrari, 1997). Resnick (1987) argues that higher order thinking skills involve complex, non-systematic, distinctive judgements, and the considerations of multiple sources and solutions.

The learning models and approaches the revised EFL textbooks employed in the Specialized Secondary Programme were supposed to shift the EFL classroom dynamics from teacher-centered to learner-centered pedagogical approaches. According to Dewey (1916), such approaches emphasize the importance of learners' independence and responsibility for the learning process, and ascribe greater importance to the learners' experience and knowledge in the classroom. In addition, the student is perceived to be an active enquirer, rather than a passive receiver of information; in other words, the learning process is a "shared activity" in which "the teacher is a learner, and the learner is, without knowing it, a teacher" (p.160).

With the change of the secondary schooling system in Libya from the academic year 2013/2014, the EFL curriculum and material for the secondary stage level was further revised. A detailed description of the secondary school system, which is the focus of this study, is presented in the following section. The document analysis of the prevailing EFL curriculum and concomitant classroom pedagogy is presented in Chapter Six.

## **2.5. The Secondary School System and Its Development: 1960-2017**

Secondary level schooling within Libya has witnessed considerable changes, as part of education reform since 1960. The General Secondary Programme, which was the only secondary level programme in Libya between 1960 and 1987 (Asker, 2012), had two components: Arts and Social Sciences (referred to as the literary section) and Natural Sciences (referred to as the scientific

section). Each section covers a wide range of subjects, with the scientific section including pure mathematics, mechanics, statistics, chemistry, physics, biology, English, Arabic studies and Islamic studies, while the literary section subjects include social studies, history, geography, English, Arabic studies, Islamic studies, Arabic literature, phonology, semantics and psychology.

In 1992, another secondary level programme known as the Specialized Secondary Programme was implemented alongside with the existing programme (Asker, 2012). The new curricula consisted of a four-year programme in which students would study the regular general subjects in the first year, and after successfully completing the first year they would specialise in one of the six majors of specialised study (Ali, 2008) (see Table 2.1). As documented in the Ministry of Education's policy statement (1992), the new programme was implemented in order to:

1. Augment the effectiveness of secondary level learning through directing the efforts and resources to specific academic domains of student interest;
2. Provide students with the necessary knowledge and skills in order to excel at tertiary education or vocational training; and
3. Meet the regional and academic needs of the country.

Each of the six different specialized domains has its own specific subjects, along with English being taught as a foreign language. With each major, Libyan students receive four classes of English every week (Ali, 2008), while teachers within the Specialized Secondary Programme would have 8 to 12 classes per week (Asker, 2012). During the first year of the programme, students study English language content tailored towards general language usage with no specific focus on students' linguistic content knowledge of their chosen specialization (Abubaker, 2017).

Table 2.1

*Secondary Schools Majors* (Source: Ali, 2008)

| <b>Secondary Schools Specialisations</b> | <b>The School Subject Areas</b>  |
|--|--|
| School of Life Sciences                  | Medical Sciences/ Agricultural Science   |
| School of Basic Sciences                 | Biology & Chemistry/ Physics & Mathematics                                       |
| School of Engineering Sciences           | Building, Electricity and Electronics, Mechanics, and Natural Resources          |
| School of Economical Science             | Administrative Sciences, Financial Sciences & Banking Information and Statistics |
| School of Social Sciences                | Religious Sciences, Arabic Language, English language, Social Sciences           |
| School of Art & Media                    | Fine & Practical Arts, Media Arts  |

However, in the following years the EFL textbook content was tailored towards the specific specialization to help build EFL students' language competence in their major, and prepare them for a level of specialization necessary for tertiary education. In other words, in the EFL textbooks in the English language specialization major, students study the following subjects: listening, speaking, phonetics, reading, writing and grammar. Onaiba (2014) states that English language specialization major was an elite specialization and many parents encouraged their children to enroll in it. Students who wished to enter this major must have achieved at least 65% in the Basic Education Certificate Examination (BECE, a high-stakes exam, see Section 2.6). Table 2.2 summarises the results of the SECE language specialization major from 2006 to 2014 in the city of Misurata where the study was conducted.

Table 2.2

*Results of the Secondary Education Certificate Examination of English Language Specialization Major, 2006-2014* Source: (Assessment and Evaluation Administrative Office of Misurata, 2017)

| <b>Year</b> | <b>No. of students entering the SECE</b> | <b>No. of students who passed the SECE</b> | <b>Percentage of students who passed the SECE</b> | <b>No. of students who failed the SECE</b> | <b>Percentage of students who failed the SECE</b> |
|-------------|--|--|---|--|---|
| <b>2006</b> | 530                                      | 448  | 85%   | 82   | 15%   |
| <b>2007</b> | 745                                      | 718  | 96%   | 27   | 4%  |
| <b>2008</b> | 603                                      | 486  | 81%   | 117  | 19%   |
| <b>2009</b> | 861                                      | 735  | 85%   | 126  | 15%   |
| <b>2010</b> | 615                                      | 573  | 93%   | 42   | 7%  |
| <b>2011</b> | 433                                      | 421  | 97%   | 12   | 3%  |
| <b>2012</b> | 310                                      | 290  | 94%   | 20   | 6%  |
| <b>2013</b> | 318                                      | 300  | 94%   | 18   | 6%  |
| <b>2014</b> | 346                                      | 294  | 85%   | 52   | 15%   |

The availability of the specialty schools depended on the size and geographical location of the area, and the accessibility of resources such as teachers. In essence, students cannot always specialize in areas of their first choice if they are unable to access a school that provides their choice. Successful students are expected to continue in the same area of specialisation in tertiary education. By the 1998/1999 academic year the four-year Specialized Secondary Programme fully replaced the General Secondary Programme, and was the only active secondary schooling programme within Libya at that time (Asker, 2012). Then, in 2009, the four-year Specialized Secondary Programme became a three-year programme.

Thereafter, the three-year Specialized Secondary Programme was considered to be part of overthrown Gaddafi regime that took place in 2011. Accordingly, the existing elected government welcomed the strong calls for the reintroduction of the General Secondary

Programme and its two divisions, the literary and scientific domains (Personnel Communication with the Libyan Ministry of Education, 2016). It was further indicated by officials at the Ministry of Education that the General Secondary Programme was considered to be the optimal secondary schooling system because many rural areas had limited access to the six schools of the Specialized Secondary Programme. Consequently, this sector of society had limited opportunities for learning, and limited choices of specialization. Thus, by the 2013/2014 academic year the General Secondary Programme had fully replaced the three-year Specialized Secondary Programme and continues to be the only active secondary schooling programme within Libya.

### **2.5.2. Schools and Classes at the Secondary Level**

This brief description of secondary level classes within Libya provides context for the study. The academic year in Libya runs from September to the end of May and consists of two consecutive terms. The first term starts in September and ends in late January, and the second term starts early February and ends in late May. At the end of the first term students have a two-week mid-year holiday. Up until 2012, students used to attend school six days a week (Saturday to Thursday), but now students attend school five days a week (from Sunday to Thursday).

Every school in Libya is built with the aim of providing a learning space that is equipped with all necessary learning facilities. The number of classes within a typical secondary school can range from 12 to 30, with each class having capacity for up to 44 students, although the average is 25. The number of classes within a school vary depending on its location (Ali, 2008). Figure 2.2 illustrates a typical classroom arrangement for the Libyan context, with students sitting in rows facing the teacher's desk which is located at the front of class with the white/black board behind him/her. Each class is 45 minutes long.

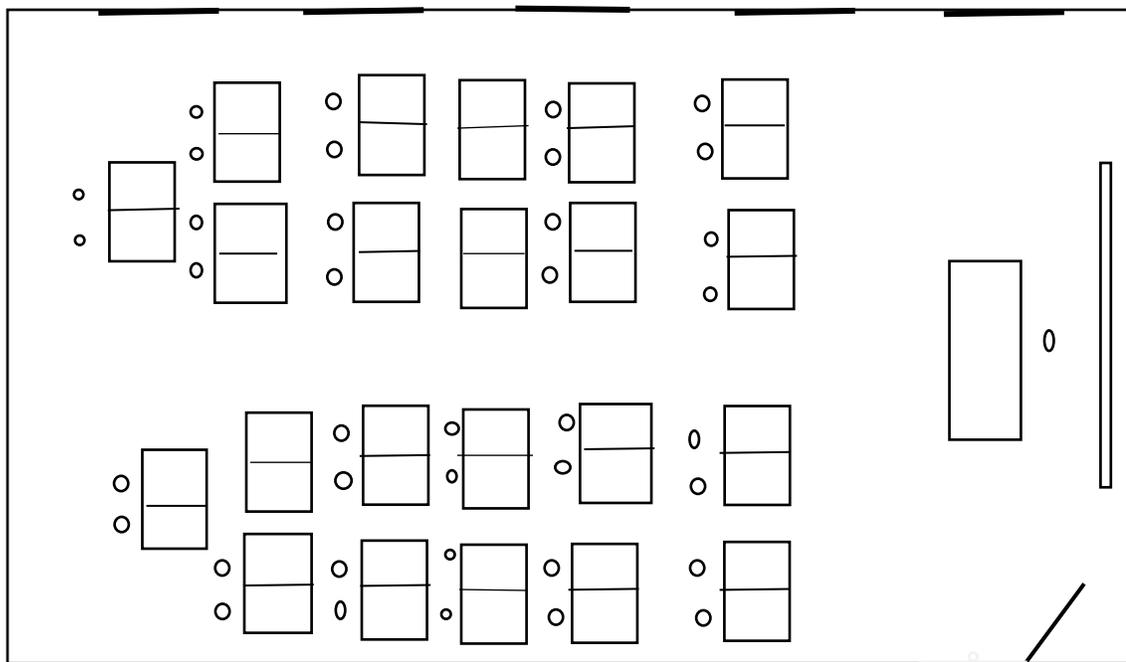


Figure 2.2: Typical Classroom Arrangement in a Libyan Secondary School  
Source: Adapted from Ali (2008).

#### 2.4.2. Characteristics of the Educational and Classroom Culture

Considering the crucial role that the culture of learning plays in shaping the classroom dynamics (Cortazzi and Jin, 1997), the Libyan culture of learning merits discussion. Thus, this section presents a description of the characteristics of the Libyan educational and classroom culture. It is generally accepted that the educational norms and practices in a society are strongly influenced by the socio-cultural dynamics of the society itself (Coleman, 1996, Tudor, 2001). The consensus is that the ‘culture of learning’ may be an influential factor with regards to what happens inside a language classroom and what is determined to be a successful language learning environment (Cortazzi & Jin, 1997). In this study, the ‘*culture of learning*’ implies that:

Behavior in language classrooms is set within taken-for-granted frameworks of expectations, attitudes, values and beliefs about what constitutes good learning, about how to teach or learn, whether and how to ask questions, what textbooks are for, and how language teaching relates to broader issues of the nature and purpose of education... Any

particular culture of learning will have its roots in the educational, and, more broadly, cultural traditions of the community or society in which it is located (Cortazzi & Jin, 1997, p.169).

It is further argued by Cortazzi and Jin (1997) that learning is inborn in social interaction and originates from cultural norms, morals and expectations that spring from the learners' instant community or from the overall society itself. Not only does society influence the culture of learning, but also the socio-economic circumstances of a particular culture play a role. It is believed that teachers and learners in a classroom may be unconscious of such culture and how it may shape the classroom practices and pedagogy. Thus, the culture of learning may be perceived as "part of a hidden curriculum" (Cortazzi & Jin, 1997, p.169). Likewise, Senior (2006) argues that both teachers and students function within a socio-cultural setting and their beliefs and expectations are influenced by the norms of that specific milieu.

In this study, belief is defined as a "proposition which may be consciously or unconsciously held, is evaluative in that it is accepted as true by the individual, and is therefore imbued with emotive commitment; further, it serves as a guide to thought and behaviour" (Borg, 2001, p.186). Both Borg (2001) and Senior (2006) emphasize the significant role that teachers' beliefs play in both teaching and life. They shape and help individuals become aware of the world, manipulate how innovative information is acquired and establish whether or not this information is acknowledged. Both students and teachers bring with them to any classroom a fixed set of beliefs and expectations about classroom etiquette, and what is to be taught and how it should be taught.

One of the main features of the Libyan educational culture is that Libyan students often take on the role of passive learners, i.e., they sit, listen and memorise the information passed on by the teacher (Deeb & Deeb 1982). In Libya, students listen as the teacher explains concepts and

respond only to direct questioning (Aldabbus, 2008). Classrooms are normally teacher-centred; teachers are perceived to be knowledgeable and have full command and competence of the subject matter being delivered to the students. It would be considered impolite to interrupt a teacher while he/she is delivering a class, or even to argue about the validity of any given information. Therefore, students show respect to their teacher by being silent and accept what they are told. Teachers are seen by students as the “suppliers of information which has to be recorded and reproduced accurately in examinations” (Alhmali, 2007, p.173). Consequently, Libyan students perceive schooling as a mandate “to receive as much information as possible which will help them move on to the next educational level” (Abubaker, 2017, p.17). Furthermore, the teacher is expected to teach to fully prepare students for exams, and to only teach what will be in the exam, i.e., ‘teach to the test’ (Abdulhamid, 2011).

Students have very little knowledge of their own to add to the education process, and no right to question the legitimacy of what they are learning (Alhmali, 2007). “The result of accepting such beliefs about teacher authority is an unacceptably passive and unequal role in learning for students, who are left with very limited opportunities for creative expression in the classroom” (O’Dwyer, 2006, p.3). Looking at this context from a student-centred learning perspective, students have few, if any, opportunities to become analytical, critical, or autonomous learners (Abdulhamid, 2011). Overall, student-classroom interaction with the delivered content is not encouraged within the Libyan context. There is no space or time given in the classroom for student to work in groups and discuss meaning. In essence, they are not given a “chance to broaden their perspectives and sharpen their understandings as they compare their ideas with others and make meaning” of the delivered material (Abubaker, 2017, p.18)

Because of the traditional role that a teacher has to adopt in a Libyan classroom, the ELT education system in Libya concentrates on the teachers' knowledge of the language and not on how to teach it. The focus is purely on raising the awareness of the semantic, syntactic and phonological knowledge of English. It is argued that by augmenting teachers' knowledge of English, they will be fully acquainted with and competent in it, and thus, would be in a strong position for delivering the acquired knowledge to language learners. In addition, the pressure to provide only accurate information encourages teachers to develop their linguistic knowledge rather than their classroom practices. It might even prevent teachers from implementing new approaches to language teaching, in order to not lose face and threaten their sense of security (Orafi, 2008).

Libya's classroom instruction (including ELT) is test-driven; it is designed to train learners for high-stakes exams and university access tests, and students' "learning depends on how well they are able to perform in their examinations" (Abubaker, 2017, p.17). These exams test rote memorisation of textbook content (Al-Buseifl, 2003). Thus, memorisation of facts and information holds prominence within the Libyan educational culture. Memory within this educational perspective is conceived as a container, in which knowledge and information is put, and retrieved whenever it is needed. Knowledge is discussed as if it is something that can be delivered or transferred to students; in other words, it is as if the students' heads are opened and the information is poured into them, instead of helping students to develop their own explanations, to reason, and to draw conclusions (Abdulhamid, 2011). In 2004, the Ministry of Education issued a report that highlights the dominance of memorisation skills within the Libyan education system and how it affects and impedes innovation and reform:

Education in Libya has a traditional character in methods and schemes. It is interested to supply students with information, but it does not care much for scientific thinking methods. Undoubtedly, the assurance on information learning by heart, for which the learner is awarded with high grades, is one of the obstacles of innovative thinking, and preparing students to knowledge production (Libyan National Commission for Education, 2004, p.65).

There is also a sustained belief that anyone can accomplish success in language learning and acquire an additional language just by conscientious, hard work, even at university level (Abdulhamid, 2011). Thus, it may be argued that the Libyan culture perceives education as a process of conveying information. A teacher's inability to answer a student's question may be perceived as an inadequacy of the teacher. Teachers and textbooks are considered to be the main source of information, and therefore fundamental components of the Libyan educational culture. "[T]extbooks represent the syllabus and dictate what should be taught in the classrooms" and [t]eachers teach according to textbooks...and achievement tests are designed based on the content of textbook" (Wang, 2006, p.50). Arguably within the Libyan educational context "textbooks used in classrooms *are* the curriculum" (Richards, 1998, p.125). Students are given textbooks for each subject, and are expected to both comprehend and memorise what is articulated in these textbooks to pass tests (Abdulhamid, 2011).

The educational learning setting within a Libyan classroom is competitive rather than cooperative. A number of studies show that learning in a cooperative working context may result in "higher individual achievement than do competitive or individualistic efforts", if properly implemented (Johnson, Johnson & Stanne, 2000, p.13). Johnson et al. (2000) report a cooperative style of learning may promote: (a) higher-level reasoning; (b) retention; (c) time on

task; (d) transfer of learning, (e) achievement; motivation; (f) intrinsic motivation; and (g) social and cognitive development, and moral reasoning. Nevertheless, Libyan students tend to compete with one another to achieve high marks, which accrue high status within Libyan society. Libyan parents wish to get good teachers for their offspring; they push them to study hard to pass the two-national high-stakes exams (Abdulhamid, 2011).

These classroom norms and dynamics may have been shaped by other dominating factors customary in the Libyan culture. Deeb and Deeb (1982) point out that prevailing Libyan cultural customs are based on Islamic moralities and regulation, and thus Libya can be viewed as a practicing Islamic society. It is common in Libya for young children to be nurtured to respect their elders, to listen conscientiously, not to argue with elders, and not to participate in adult family discussions. If a child does so, he or she will be disciplined by their parents (Abdulhamid, 2011).

In summary, it may be argued that both the religious and socio-cultural norms of the Libyan culture, such as respect for elders, the avoidance of interruptions of and arguments with elders, have informed traditional teacher-centred Libyan classroom practices. Having reviewed the evolution of the Libyan education system and the characteristics of the Libyan educational culture, the next section discusses ELT in Libya.

### **2.4.3. Secondary Level EFL Teachers**

The teaching profession is very popular among women in Libya (Onaibia, 2014; Shihiba, 2011). EFL teachers in secondary level education in Libya must hold either a teaching diploma or a Bachelor Degree in the Arts and Social Sciences, or Education. However, graduates holding an English degree from a Department of English at a Faculty of Arts and Social Sciences have had no theoretical or practical training in EFL learning theories and teaching (Onaibia, 2014;

Shihiba, 2011). In contrast, graduates from the Faculty of Education receive considerable theoretical foundational knowledge and practical experience about EFL teaching (Onaibia, 2014; Shihiba, 2011). At the Faculty of Education, students in the Department of English are also exposed to theories of psychology (such as general psychology, psychology and development, and children's health) and their application within the domain of education (Shihiba, 2011). In addition, students are enrolled in a four-week teacher training programme where students shadow EFL teachers in their classrooms. In teacher-training programmes students are also given the opportunity to put their teaching knowledge into practice by teaching Basic or Secondary EFL classes.

The teacher-training programme can be described as an apprenticeship. In simple terms, an apprenticeship is a “process through which a more experienced person assists a less experienced one, providing support and examples, so the less experienced person gains new knowledge and skills” (Dennen & Burner, 2007, p.426). An example of an apprenticeship in our daily lives would be a parent teaching a child how to eat, and the process by which a person may learn to be a carpenter or plumber. In these examples, one would not expect the learner to learn just by observing a single demonstration. Instead, the parent or experienced other conveys the necessary new knowledge to the novice in small tasks. Plus, guidance is offered so that the tasks undertaken by the learner are within the reach of the learner's current ability level, i.e., within the Zone of Proximal Development (ZPD, Vygotsky, 1986). Vygotsky coined the term ZPD to help explain the way social and participatory learning takes place. ZPD is:

the difference between the mental age, or the level of the actual development, which is defined by the problems that can be solved independently, and the level, which is reached

by the [person] through the solution of problems accomplished not independently but in collaboration. (Vygotsky, 1991/2014, pp.399-400)

It is worth noting that the educational advantage of apprenticeship is not only limited to the apprentice learning psychomotor or professional skills, but also, following Vygotsky, extends to the development of cognitive and metacognitive learning processes (Dennen & Bruner, 2007). In addition, apprenticeship learning may occur in both formal and informal learning environments. Vygotsky coined the concept of ZPD as a way for psychologists and educators to think about children's development and how they learn, and develop problem-solving abilities required to perform a range of developmentally related tasks (Ormrod, 1995).

The Libyan education system assumes that EFL teachers working at the secondary level have sufficient competence and knowledge to “develop students’ abilities and prepare them” for higher education (Ali, 2008, p.14). Ali (2008) further argues that secondary level teachers are also expected to take on the role of “facilitators and managers of learning” and to be able to “engage their students in the learning process” (p.14). Although, Libyan EFL teachers have welcomed the contemporary models of language teaching within the Libyan context, they were “not translated into classroom practices” (Orafi & Borg, 2009, p.249). The challenges that secondary level EFL teachers face in their classroom is developed in more detail below.

#### **2.4.4. Challenges of Teaching English at the Secondary Level**

After the Specialized Secondary Programme became a three-year one, the Libyan Ministry of Education documented that there was considerable shortage of qualified EFL teachers, particularly at the secondary level stage. In 2004, the Libyan National Commission for Education, Culture and Science highlighted several challenges that Libyan secondary EFL teachers faced in their day-to-day teaching, including:

1. The majority of teachers have limited information on how to employ communicative language approaches in their classrooms;
2. Many schools face a lack of teaching resources such as overhead projectors, language laboratories and in some cases even tape recorders;
3. Most of the time, students do not have the opportunity to interact inside the classroom, either because of the time constraints, or overcrowded classrooms;
4. Teachers are suffering from a shortage of in-service training to assimilate up-to-date changes and research development in teaching and learning foreign languages; and

On the matter of the shortage of qualified Libyan EFL teachers many Libyan researchers including Ali (2008) and Omar (2014), report that Libyan EFL teachers lack the necessary proficiency to be teachers. Omar (2014) further notes that the revised EFL curriculum requires knowledge about the language, which Libyan EFL teachers lacked and to deliver such knowledge to students may have even be problematic.

Moreover, Orafi and Borg (2009) report that teachers altered communicative activities, such as pair-work, to teacher-centered actives that involve teacher asking questions and students answering them. They also emphasise the considerable use of translation and L1. These findings were initially reported by Orafi (2008) who investigated the extent to which Libyan EFL teachers implemented the revised curriculum in their language classrooms. Orafi (2008) concludes that there was a very weak degree of alignment between the set curriculum and the language instruction implemented by Libyan EFL teachers. In essence, Libyan EFL teachers did not implement the innovative practices that the Ministry of Education had expected them to do. In line with Orafi (2008), Omar (2014), who assessed Libyan EFL teachers' accounts of their language classrooms, reports that the traditional grammar translation method was the most

employed method of EFL teaching by Libyan teachers. Orafi (2008) argues that the lack of implementation of the set curriculum may be distrusted by teachers because of the following factors:

1. Teachers' beliefs, assumptions, and knowledge (Woods, 1996) about language teaching and learning;
2. Teachers' prior language learning experience;
3. The prevailing Libyan culture of learning;
4. The governing examination system; and
5. Limited opportunities for teacher development.

In addition, Altaieb's (2013) doctoral research argued other factors which hindered Libyan EFL teachers' implementation of the set communicative language teaching (CLT) curriculum and practices.

1. Limited time for teaching CLT materials;
2. Insufficient funding;
3. Students' limited language skills;
4. Teachers' lack of training in CLT;
5. Large class sizes;
6. Lack of support from colleagues and administrators;
7. The focus on rote memorization in teaching and learning;
8. Students' lack of motivation for developing communicative competence; and
9. Students' resistance to class participation.

According to Altaieb (2013), Orafi (2008) and Orafi and Borg (2009), it can be argued that "large scale reform took place only on the textbook level and ignored other components vital to

the success of curricular reform”, such as technical facilities, teachers’ professional development programmes, teachers’ beliefs, students’ needs, school space, and time (Altaieb, 2013, p.ii). In addition, the implementation of the desired curriculum necessitates the availability of resources that promote students’ participation through dialogue, playing games, role-play and problem-solving activities (Shihiba, 2011).

Even after a decade of reform within the Libyan education system, these challenges continue to prevail. Omar’s (2014) doctoral research reports that Libyan schools lack facilities such as visual aids and technical resources that can support language teaching. Omar’s (2014) findings also highlight the challenges posed by textbooks availability and a number of contextual factors. Teachers were provided with incomplete packages, each language teacher is supposed to receive a student’s course book, student’s workbook, teacher’s guide, and CDs or cassettes. However, Omar (2014) reports that not all teachers received the teacher’s guide, and the CDs/cassettes were not provided at all. Contextual factors such as lack of collaboration between schools and parents, and lack of teacher training programmes were documented as contributing factors that brought about challenges for Libyan EFL language teachers. Therefore, it can be concluded that the “mismatch between the realities of the classroom... and the principles and goals of the new curriculum created a significant challenge for teachers” (Altaieb, 2013, p.iii). Thus, there is dissonance between the “what is expected in the new curriculum and what is actually being done in classrooms” (Altaieb, 2013, p. iii). Consequently, these challenges may have encouraged many teachers to be involved in the Libyan Study Abroad Scholarship Programme (LSASP - Abdulhamid, 2011). In essence, the high number of teacher enrolment in the LSASP may suggest that Libyan EFL teachers are anxious about the quality of their classroom rituals and interested in acquiring innovative teaching approaches that may improve the quality of their

students' learning. In addition, the increase in financial support from the Libyan government highlights the fact that change is a desired objective for the Libyan nation as well as the Libyan teachers. The LSASP which dates to the 1970s was described by Abdulhamid (2011) as a teacher development or enhancement programme, as it aims to develop "the capacity of Libyan citizens [for] addressing the needs of Libyan public institutions, academia and Libyan society as a whole through the pursuit of graduate and post-graduate studies"<sup>14</sup>.

## **2.6. The Role of Examinations in Libya**

Within the Libyan education system examinations play a fundamental role, and for decades, have been the prevalent form of student assessment and evaluation. It can be safely said that examinations "legitimate the education enterprise" in Libya. Libya has a very "long tradition of education that is reflected in individual success in standardised examinations" (Chapman, & Snyder, 2000, p. 462). When students enter their fourth year of primary schooling they face various examinations. As a result, high-stakes testing has become a "natural part of school life and for students it is part of every day schooling" (Minarechová, 2012, p.87). For decades and even after the overthrow of the Gaddafi regime, Libyan students have sat and continue to sit two fundamental examinations the BECE and SECE. Each of these examinations has two purposes: (1) to determine the eligibility for a school graduates' certificate for that level, and (2) to determine a students' eligibility and placement in the educational options at the next level. At these two stages, Libyan students are required to demonstrate their learning of subject matter in high-stakes, external exams that are administrated by the Libyan Ministry of Education. The BECE examination certifies the completion of lower-secondary school studies (Grade 9), while

---

<sup>14</sup> Libyan-North American Scholarship program <http://www.cbie.ca/data/libya/Home/default.htm>

the SECE certifies the completion of secondary school studies (Grade 12). If the student passes the exam they move on to the next grade level.

Furthermore, the SECE result determines if students graduate from high school and to which university or college they will be admitted. Hence, these are “events that may have a major influence on students’ lives” (Marchant, 2004, p.3). If the student fails the exam, the only two options available are to re-sit the exam or repeat the grade (if the re-sit is failed). It is important to note that despite the emphasis in the EFL Teacher’s Book on the importance of implementing communicative approaches, teachers are not provided with methods of evaluation and assessment. Instead, they are provided with rigid guidelines for both pedagogical approaches and assessment from the Ministry of Education that require “teachers to adhere to a systematic and extensive schedule of formal assessment based on calculated examination and test marks” (Askar, 2012, p.22). ELT examinations in the past had the tendency to test rote memorisation of vocabulary and explicit grammar rules, and little attention was paid to either listening or speaking skills (Al-Buseifl, 2003). In line with Al-Buseifl (2003), Alhmali (2007) condemns the high-stakes examination system, particularly the SECE, because of its emphasis on the “rote recall of information” and for ensuring “great power over the learners at key times of the year” (p.ii). Whether the rSECEE still echoes the importance of memorization, and has the tendency to test rote memorisation is confirmed through empirical evidence, and thus Phase I investigates to what degree the rSECEE is aligned with Libya’s EFL content standards.

As part of the reform policy introduced in 2009, the BECE and SECE methods of examinations were changed. Originally, both BECE and SECE were hybrid in nature consisting of open ended- constructed response and discrete-point item questions (Orafi & Borg, 2009). As documented by Onabia (2014), the former version of BECE of English test, which is very similar

in test format to SECEE, made use of textbook-based questions covering a range of language skills reading, vocabulary, grammar and writing. The exam questions were randomly presented with no clear division between the tested constructs. Students answered questions on the same question paper, and were marked by EFL teachers and inspectors. The BECE's of English reading text was either a short story or a non-fiction piece with several WH-questions that student could answer by "lifting sentences" from the given text (Onaiba, 2014, p.19).

The former examination systems evaluating English as a subject have been criticized by researchers, including Al-Buseifl (2003), Alhamli (2007), Onaiba (2006), Orafi (2008), Orafi and Borg (2009). They argue that the former system tested rote memorization of vocabulary and grammar rules and students' ability to recall of information, rather than their ability to integrate and produce (Al-Buseifl, 2003, Alhamli, 2007; Onaiba, 2006). During the school year students "were encouraged to memorize huge chunks of material and then regurgitate them" in the final examination (Onaiba, 2014, p.20). Onaiba (2014) further questioned the validity of the former examinations, as the set test items did not appear to measure what they were supposed to measure, and listening and speaking skills were not tested. Although the BECE and SECE curriculum is communicatively-oriented, not testing either listening or speaking is liable to have encouraged teachers to ignore them in their day-to-day classroom instruction (Onaiba, 2014). In essence, there was a lack of alignment between assessment and curricula content, which may, in turn, have had disadvantageous effects on the education system (Hughes, 2003; Linn, 2000; Shohamy, 1997; Tan, 2008; Wall, 2005). Alignment in this sense can be defined as the degree of agreement among curriculum standards, instruction and assessments (Cheng & Fox, 2017). Moreover, the former examination system was criticized because of the degree of cheating. According to Onaiba, (2014) student-to-student cheating or teacher-to-student cheating was

possible because “students of the same grade were tested on the same subjects with the same exam papers and sat the exam quite close to one-another” (p.21).

Consequently, there was a call for former examination systems to be replaced by more comprehensive and sound assessment tools that consider the set EFL curricula objectives (Alhmali, 2007; Onaiba, 2006; Orafi, 2008). Thus, in 2009 a reform policy for the Libyan examination system was implemented. As noted by the Ministry of Education the revised testing system and exams, i.e., automated scoring assessment system, “aim to develop ways and methods of examinations that are in accordance with modern scientific developments”, and computers were employed “to monitor grades, issue certificates” and to help “students review their results” (Ministry of Education, 2008, p.11). The aim of the Libyan Ministry of Education was to simplify the way in which examinees answer exam questions, align exam content with curricula content, provide students with accurate measurement of their performance, minimise the chance of cheating, and technically mark answers to help issue results with more efficiency (Ministry of Education, Decree No. 6, 2007). In essence, the Ministry of Education has put efforts into reducing the possibility for Libyan students to attain marks by fraudulent means, and thus enhancing the face validity of the revised testing system.

The comparison of the rSECEE with the former version (see Appendix A) clearly shows a difference between the two versions in terms of format and style (Onaiba, 2014). Table 2.3 summarises the differences between the rSECEE (from 2009 until 2018) and the former SECEE (prior to 2009).

It is worthy of note that although the rSECEE changed in terms of test content and format, it did not change items of its role, i.e., to date the revised is a high-stakes test that makes life changing decisions for Libyan students. The revised examination was first introduced for both

BECEE and SECEE in the 2008/2009 academic year. Therefore, when conducting the primary data collection for the research, the rSECEE had been in practice for nine years. However, the revised General Secondary Programme and curriculum had only been in place for three years. Onaiba (2014) was the first to conduct a study about the washback effect of a revised BECEE on teachers' instructional practices, materials and curriculum.

To the best of my knowledge, no researcher has examined how the change in the high-stakes SECEE has influenced the Libyan EFL language classroom. Thus, this adds to the study's originality and importance. Bearing in mind the key role the SECE plays in the Libyan context and with it being the principal focus of this research, the SECEE merits discussion. Thus, the next section contains a description of the rSECEE.

Table 2.3

*Differences between the rSECEE and the Original SECEE*

| <b>rSECEE (As of 2009)</b>   | <b>Former SECEE (Prior to 2009)</b>   |
|--|---|
| <ul style="list-style-type: none"> <li>• Students transfer answers to mechanically scored answering sheets</li> <li>• Discrete-point item (such as multiple choice, true-false, and matching items)</li> <li>• Exam instructions and means of transferring answers to answer sheet are given in Arabic</li> <li>• Questions are written in both Arabic and English</li> <li>• Objectively scored</li> <li>• Question papers with individual candidate's name, student number and school name and city are provided by the Ministry of Education</li> </ul> | <ul style="list-style-type: none"> <li>• Students provide their answers in the same question booklet.</li> <li>• Open-ended-constructed response and discrete-point item questions</li> <li>• Exam instructions are given only in English</li> <li>• Questions are only given in English</li> <li>• Subjectively and objectively scored</li> <li>• Students individually write their own names and student numbers on the administered exam papers</li> </ul> |

### 2.6.2. The SECEE

For the SECE certification, the students must take all subject matter tests that are assigned in their domain. The focus of the research presented in this dissertation is on the rSECEE for the literacy division, which is the one of compulsory tests for university entrance. Similar to the rBECCE, the rSECEE is a pencil-and-paper exam that takes place twice a year. Selected experienced inspectors construct two versions of the rSECEE each year. One exam is chosen for the first sitting, and the other for the re-sit. Governed by a set of regulations and test specifications stemming from the subject matter objectives, the test constructors ensure that both versions of the rSECEE are identical in terms of length, scope and difficulty (Onaiba, 2014). From my understanding, the creation of rSECEE may have been drawn from an item bank that was adopted for the whole test development process.

Libyan students entering either the BECE or SECE are required to take two mid-term examinations, which account for 40% of their final grades. Some high achievers enter the BECE or SECE with a full 40% for almost all subjects. As with other standardized tests around the world the rSECEE has a set rules formulated by the Ministry of Education, such that every examinee receives the same directions and has the same time restrictions and resources. Within the rSECEE testing context, each subject is tested through a set of multiple-choice items that are administrated by the teachers, and each student works independently on machine-scored answer sheets. The students are given three hours to complete the test. In contrast to the former SECEE, the revised version follows a discrete response item testing style with 60 items: 25 true/false items; 25 multiple choice (MC) items; and 10 items matching the word to its definition (see Appendix B for a sample). The MC item format is the mostly commonly used format within high-stakes testing contexts because of its “many positive psychometric characteristics, its long

history of research evidence, its versatility in testing most cognitive knowledge, its relative (apparent) ease to write, store, administer, and score” (Downing, 2002, p.235).

It can be argued that the stakes of the SECE are visible to Libyan society as top achievers and successful secondary students are awarded with scholarship programmes to study abroad to gain their undergraduate and graduate degrees. The consequence of reward associated with SECE students’ scores have “increased the value of earning top grades among [Libyan] students and has arguably influenced their motivation to work alone and increased competitiveness in their classrooms” (Askar, 2012, p.22). Tables 2.4 and 2.5 summarise the results of the SECE from 2015 to 2017 for the city where the study was conducted.

Importantly, to the best of my knowledge, the Libyan Ministry of Education has not issued an official mandate aligning national mandated testing with the national content standards and students’ academic performance. In addition, it is not obligatory for cities or schools to provide coherent information regarding students’ attainment of the articulated standards.

Given that this study explores the relationship between the degree of alignment and the washback of the rSECEE, in Chapters Three and Four, a review of the literature on alignment, high-stakes testing, and washback is provided as background.

Table 2.4:

*Results of the SECE Scientific Stream, 2015 – 2017*(Assessment and Evaluation Administrative Office of Misurata, 2017).

| Year | SECE FIRST SIT                    |                                     |                                     |                                   | SECE SECOND SIT                   |                                     |                                     |                                   |
|------|-----------------------------------|-------------------------------------|-------------------------------------|-----------------------------------|-----------------------------------|-------------------------------------|-------------------------------------|-----------------------------------|
|      | No. of students entering the SECE | No. of students who passed the SECE | No. of students who failed the SECE | Percentage of students who passed | No. of students entering the SECE | No. of students who passed the SECE | No. of students who failed the SECE | Percentage of students who failed |
| 2015 | 2732                              | 2522                                | 210                                 | 92%                               | 210                               | 144                                 | 66                                  | 69%                               |
| 2016 | 2633                              | 2211                                | 422                                 | 84%                               | 422                               | 401                                 | 21                                  | 95%                               |
| 2017 | 2793                              | 1933                                | 860                                 | 69%                               | 860                               | 566                                 | 294                                 | 66%                               |

Table 2.5

*Results of the SECE Literacy Stream, 2015 - 2017*(Assessment and Evaluation Administrative Office of Misurata, 2017).

| Year | SECE FIRST SIT                    |                                     |                                     |                                   | SECE SECOND SIT                   |                                     |                                     |                                   |
|------|-----------------------------------|-------------------------------------|-------------------------------------|-----------------------------------|-----------------------------------|-------------------------------------|-------------------------------------|-----------------------------------|
|      | No. of students entering the SECE | No. of students who passed the SECE | No. of students who failed the SECE | Percentage of students who passed | No. of students entering the SECE | No. of students who passed the SECE | No. of students who failed the SECE | Percentage of students who failed |
| 2015 | 1227                              | 1069                                | 158                                 | 87%                               | 158                               | 122                                 | 36                                  | 77%                               |
| 2016 | 1267                              | 1020                                | 247                                 | 81%                               | 247                               | 233                                 | 14                                  | 94%                               |
| 2017 | 1090                              | 881                                 | 209                                 | 81%                               | 209                               | 136                                 | 73                                  | 65%                               |

## **Chapter III**

### **Literature Review**

Näsström (2008) argues that standards, teaching and testing constitute the three fundamental components of an education system. The definition and functions of these components are discussed along with the triangular relationship between standards, testing, and teaching and learning. Finally, this chapter presents an overview of alignment, in which a more detailed definition of alignment and its contemporary research approaches is offered.

#### **3.1. The Relationship Between Standards, Testing, and Instructional Activities and Materials**

##### **3.1.1. Standards**

The concept of standards within different educational contexts, different countries, and at different periods in history has had various meanings (Näsström, 2008). However, standards are presently understood as objectives, learning outcomes, benchmarks, or curriculum standards (Anderson, 2002; Popham, 2003). Almost all education systems have standards, which are statements that define the “goals for student learning and focus the attention of teachers, students, parents and all the others concerned with education” on what students need to know, understand, and be able to do (Bertenthal & Wilson, 2005, p.2). In other words, standards are the ‘benchmarks of aspiration’ and, as argued by Smith (2005), standards may be seen as the coin of the educational realm because they form the framework for the curriculum in which central ideas, concepts and skills are stated. It should be emphasized that standards are not the curriculum, rather they are the foundations for the curriculum. Standards also serve as the basis for selecting textbooks, setting instructional priorities, developing assessments, and identifying

what information is to be acknowledged as evidence that students have attained the standards (Bertenthal & Wilson, 2005). In an ideal world, the goal of an educational process is to help students achieve the target standards, and the process of assessment aims to measure students' achievements against the standards set (Näsström & Henriksson, 2008; Biggs, 2003; Biggs & Tang, 2003). Moreover, standards ought to be clear and thorough, reasonable in coverage, technically correct, built around a conceptual framework reflecting a high level of intellectual engagement, and must rigorously describe examples of performance expectations for students in order that everyone involved in the process knows what is expected of them (Bertenthal & Wilson, 2005; La Marca et al., 2000).

It is important at this stage to differentiate between *content* and *performance* standards. The former, according to Popham (2003), are broad descriptions of what students are supposed to both know and be able to do, whereas, the latter are descriptions of the level the students should attain of the set knowledge and skills. Therefore, content standards within an education system can be said to define the 'what' and performance standards define the degree of sophistication of the 'what'. For the purpose of this study, I employed the term 'standards' to refer only to content standards, as the documents obtained for the Libyan Ministry of Education only define what students are supposed to both know and be able to do at secondary level education; hence, content standards.

### **3.1.2. Assessment**

The term 'assessment', which ranges from classroom observation to a high-stakes national test, is a systematic procedure for collecting information regarding students' achievement (Bertenthal & Wilson, 2005). Furthermore, assessment serves the needs of, and provides critical information for, nearly all parts of an education system, including "guiding instructional

decisions, holding schools accountable for meeting learning goals, and monitoring program effectiveness” (Bertenthal & Wilson, 2005, p.4). Assessment may also be used by authorities (including teachers) as means for demonstrating their objectives for student learning (Bertenthal & Wilson, 2005; Smith, 2005). Furthermore, assessment results are expected to offer accurate information to policy-makers, educators, students, and parents about students’ performance (Herman, Webb, & Zuniga, 2007). In determining the scope of assessment, it should be recognised that “[t]esting cannot be neutral on what is taught and learned. Any test is an expression of values on teaching and learning” (Cole, 1999, p.1).

It is important to note that although assessment can serve many purposes, no single assessment serves all purposes. To support valid inferences, every single assessment has to be designed exclusively to achieve its purpose (Bertenthal & Wilson, 2005; Cizek, 2001). For example, an assessment that is developed to provide information about students’ problems with a particular theoretical concept in order to inform future classroom practices would be designed differently from an assessment that is to evaluate the effectiveness of a new policy. The former would necessitate a rigorous and thorough testing of students’ understanding of a particular theoretical concept, whereas the latter involves a much broader form of assessment which covers all the topics that are considered important by the reform policy. Furthermore, results from either of these assessments would not validate the other (Bertenthal & Wilson, 2005).

Although some researchers use the terms ‘examination’ and ‘test’ interchangeably, others make a clear distinction between the two. According to Vernon (1964) an examination is “devised to assess the attainment and skills of pupils or students in a particular subject, whether by objective-type or by conventional written, oral or practical questions. All questions refer to a syllabus which has been defined by a teacher or examiner” (p.2). By contrast, a test is a

published instrument that has been developed by people officially trained in mental testing and statistical methods. Research Association [AREA], American Psychological Association [APA], and National Council of Measurement in Education [NCME] (2014) necessitate the piloting of test items and the adhering to the set standards which can then “enable the tester to interpret how far a pupil’s score or mark is superior or inferior to those of other similar pupils” (Vernon, 1964, p.2). Based on this distinction, for many people, including the researcher, the word ‘test’ evokes images of traditional, paper-and-pencil forms including multiple choice questions and true-false quizzes (Popham, 2003). This provides a rationale for why a growing number of educators and researchers prefer to use the term ‘assessment’, which includes both formal and informal (such as observations and portfolio assessment) forms of testing. In an ideal education system, assessment should be considered in the context in which it functions, because they are factors other than the test itself that may affect the type and extent of the impact of assessment in educational contexts (Burrows, 2004; Cheng, 2004; Cheng et al., 2015; Spratte, 2005).

Assessment is one of a number of components, including instruction, curriculum, professional development, and resources, that interact in the classroom and school to support student learning (Bertenthal & Wilson, 2005; Cumming, 2009; Davison & Leung, 2009; Popham, 2003; Tan & Turner, 2011; Turner, 2009, 2012). For assessment to serve its function it must be closely linked to objectives and instruction in order that all three elements are directed towards the same end. In this context, assessment is not seen as something separate from teaching and learning, but as “an integral part of good pedagogy, which has the potential to improve student achievement” (Masters, 2005, p.30). In other words, assessment ought to measure what has been taught to students, and what is taught should reflect the objectives for student learning emphasized in the standards (Bertenthal & Wilson, 2005; Cumming, 2009;

Popham, 2003). Consequently, all the components in an education system have to be “built on a shared vision of what is important for students to know and understand” about any subject matter, “how instruction affects that knowledge and understanding over time, and what can be taken as evidence that learning has occurred” (Bertenthal & Wilson, 2005, p.4).

### **3.1.3. Teaching**

According to Popham (2003) teaching is the third fundamental component of any education system. It is an inclusive term that includes all instructional opportunities that promote actual student learning, and the activities which teachers carry out to assist their students in attaining the standards and curriculum objectives. In essence, teachers are responsible for what students learn. They determine how much is to be taught, at what pace, and with what materials. Consequently, learning is driven by what teachers and students do in the classroom. In other words, teachers as be said to be the curriculum makers (Connelly, Clandinin, & Fullan, 1993). Besides, teachers have to manage complicated and demanding situations, channeling the personal, emotional, and social pressures of a large group of students in order to help them not only learn the subject matter content, but also become lifelong learners (Black & William, 2010). Moreover, Palmer (1998) stresses that good teaching is not simply a technique. Instead, good teaching “comes from the identity and integrity of the teacher” (p.11).

### **3.1.4. The Triangular Relationship between Standards, Testing, and Teaching and Learning**

For the past century, there has been a growing body of research that supports an important educational truism; that what and how much students are taught is associated with, and more likely to influence, what and how much they learn (Anderson, 2002). According to Burstein and

Winters (1994), the question may no longer be ‘what students know and can do’, but rather ‘what students know and can do as a result of their educational experience’.

Figure 3.1 illustrates the relationship between three fundamental components, standards, teaching (instructional activities), and assessment of any education system. The sides of the triangle represent relationships between pairs of the components:

- Standards with assessment (side A);
  - Standards with teaching (side B); and
  - Assessment with teaching (side C).
- 

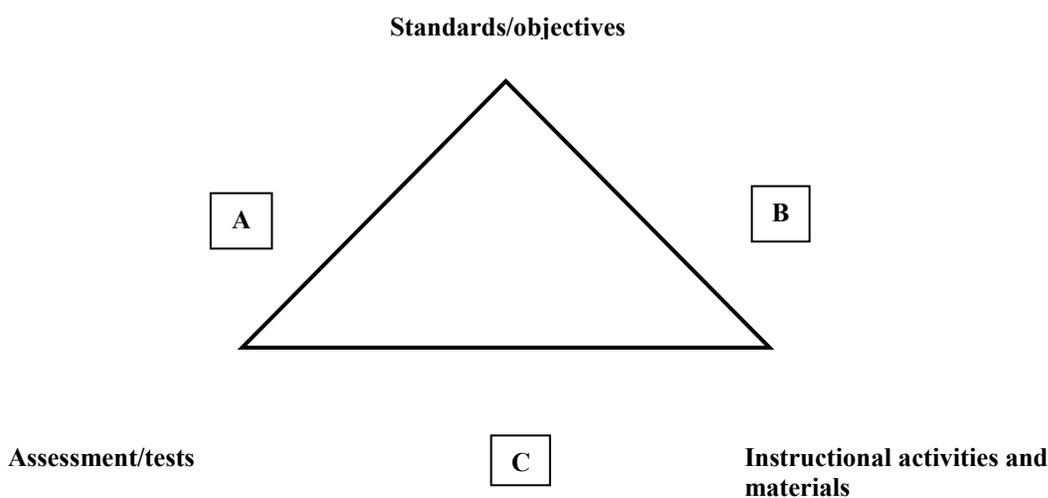


Figure 3.1: Relationship among standards, instructional activities, and assessment/tests  
Source: Anderson (2002)

The relationship between standards and assessment (side A) has been researched through a concept known as content validity, which is primarily concerned with the question ‘to what extent does the test measure the fundamental curricular objectives?’ (Anderson, 2002). Studies

that have investigated content validity include Buckendahl, Plake, Impara and Irwin (2000), Kendall (1999), and Sireci, Robin, Meara, Rogers, and Swaminathan (2000). Similarly, alignment research is concerned with the same question. Such research includes Roach, Elliott, and Webb (2005), Webb (1999, 2002), and Webb, Herman, and Webb (2007). Content validity refers to the degree to which the content of an assessment instrument assesses the assigned content or how well the content material was sampled in the measuring instrument (Brualdi, 1999; Martone & Sireci, 2009; Messick, 1989; Ruhio et al., 2003). Sireci (1998a) adds that for one to argue that a test is “valid for a particular testing purpose, it must be shown that the items and tasks composing the test are representative of the targeted content domain” (p.299). If a test is found to have high content validity, its content is judged to be compatible with the testing purpose and the subject matter concepts.

Studies that analyse the relationship between standards and teaching and teaching material (side B see Figure 3.1 above), include Buckendahl, Plake, Impara, and Irwin (2000), Ippolito (1990), and Pickreign and Capps (2000) who compare the geometry concepts used in kindergarten to Grade 6 textbooks with mathematics standards documents. Textbooks are frequently the operational means of delivering the curriculum to students (National Research Council, 2012). It is argued that textbooks may have a positive impact on teachers and classroom pedagogy during curriculum implementation (Harmer, 1991; Richards, 1998). These included: relieving teachers from the pressure of searching for material and providing a guide to teach more effectively. On the other hand, textbooks are condemned for impeding teacher development. They are seen as impeding development by:

- 1) Absolving teachers of responsibility on a day to day basis;

- 2) Leading to “the unjustifiable attributions of qualities of excellence, authority, and validity to published textbooks” (Richards, 1998, p. 131); and
- 3) Causing a “reduction of the level of cognitive skills involved in the teaching if teaching decisions are largely based on the textbook and the teacher’s manual” (Richards, 1998, p. 132).

Side C in Figure 3.1 above, the relationship between teaching and assessment is primarily researched through the lenses of content coverage and the opportunity to learn (Burstein, 1993). However, it is worth noting that there are differences between studies that research content coverage and those that research opportunities to learn. The former start by examining the instructional activities and materials and are mainly concerned with the questions ‘is what we are teaching being tested?’ or ‘is what we tested being taught?’. Such studies include Schmidt and McKnight (1995). In contrast, the latter typically start with an examination of assessment tasks and items, and are concerned with the question ‘are we teaching what is being tested?’ (Anderson, 2002; Winfield, 1993).

The whole triangle represents a relationship referred to as *alignment* or *curricular alignment*, which implies that there should be a robust connection between standards and assessment, between standards and teaching, and between assessment and teaching (Anderson, 2002; Biggs, 2002, 2003; Biggs & Tang, 2007; La Marca, Redfield, Winter, & Despriet, 2000; Martone & Sireci, 2009).

It is worth noting that both content validity and alignment research use subject-matter experts (SMEs) to evaluate test items and tasks with respect to their match to test specifications (such as strands or content areas) as well as their relevance to the domain of interest (Webb, 1999). However, current alignment research takes the evaluation one step further and examines the

congruence between items and the intended objective within a strand, and then reports the findings summarised by objective and/or by strand (Martone & Sireci, 2009). Furthermore, certain alignment research takes into account the content of instruction (i.e., what was taught to the students in the classroom). Thus, content validity, content coverage and opportunities to learn research come under the umbrella term ‘alignment research’, which is one possible means for demonstrating or assessing the link between the three components (Biggs, 2003; La Marca et al., 2000; Martone & Sireci, 2009; Webb, 1999). This level of evaluation highlights that alignment research can provide a more comprehensive view of the whole education process and is arguably an important extension to the analysis and information that typical content validity research provides (La Marca, 2001; Martone & Sireci, 2009). This argument provides the rationale for current study’s application of an alignment research approach, rather than the traditional content validity analytical approach, to investigate whether or not the rSECEE adequately measures the concepts and skill areas represented in the Libyan EFL academic standards, and whether or not the rSECEE high-stakes tests capture the meaningful aspects of student thinking and learning.

### **3.2. Overview of Alignment**

Successful schooling is based on the synchronisation of the three fundamental components of an education system: standards, assessment, and instruction (Elliott, Braden, & White, 2001; Webb, 1997, 2002; Webb, Horton, & O’Neal, 2002). The extent to which these three components work in harmony towards students’ learning is referred to as alignment, which is a complex but increasingly important topic (Bloom, Madaus, & Hastings, 1981; Impara, 2001; Tyler, 1949; Webb, 1999). Alignment between assessment and content standards has long been recognised as evidence of test validity (AERA, 1999; Impara, 2001; Resnick et al., 2003); for example, as part of the mastery-learning movement that was evident in the early 1960s,

assessment tasks were analysed to see if they were congruent with the behavioural objectives (Cohen, 1987).

### **3.2.1. Extended Review of Alignment**

As noted earlier (section 1.1) that the present study considers alignment to be “the degree to which expectations (i.e., standards) and assessment are in agreement and serve in conjunction with one another to guide the system toward students learning what they are expected to know and do” (Webb, 1997, p.4). If the three components work in harmony to deliver a consistent message about what is valued in the educational process (Webb, 1999), the students may then have a better opportunity to learn and truly demonstrate what they have attained (Anderson, 2002; Biggs, 2003; Farenga, Joyce, & Ness, 2002; Martone & Sireci, 2009). With reference to an education system delivering a consistent message, Porter (2002) states that:

[a]n instructional system is to be driven by content standards, which are translated into assessments, curriculum materials, and professional development, which are all, in turn, tightly aligned to the content standards. The hypothesis is that a coherent message of desired content will influence teachers’ decisions about what to teach, and teachers’ decisions, in turn, will translate into their instructional practice and ultimately into student learning of the desired content (p.5).

Alignment in this sense and in accordance with Näsström (2008) can be compared to the links in a chain and the three components of an education system (i.e., standards, assessment, teaching) represent the jewels that supported by and linked together by the chain (see Figure 3.2).

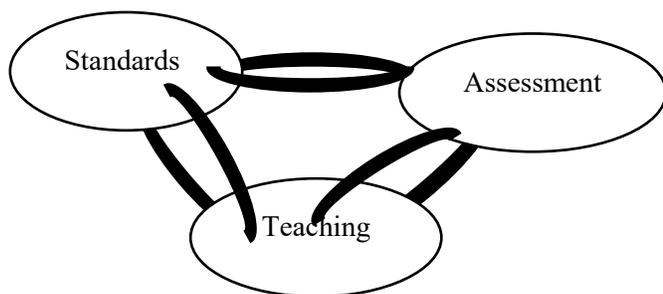


Figure 3.2: Alignment as Links between the Components of an Education System  
Source: Näsström (2008, p.20)

The strength of the chain depends on the strength of each link and determines how well the jewels are held together. For instance, if one of the links is weak it may easily break, and the chain may fall apart. Nevertheless, if all the links are strong the chain may then be able to resist any external force (Näsström, 2008). In relation to this analogy, if the links are weak, in other words if there is weak alignment, the three components may disconnect; fall away from one another and send inaccurate information to students about what they ought to know and be able to do, and may result in negative washback (Linn, 2000; Shohamy, 1997; Tan, 2008; Wall, 2005). This lack of alignment may cause teachers to ignore the set standards and instead teach only what is going to be tested, and thus the students may get less opportunity to attain the set standards (Herman, Webb, & Zuniga, 2007; Resnick et al., 2004; Winfield, 1993). Thus, negative washback may harm the teaching and learning experience and may even substantially affect students' academic performance (Frederiksen, 1984; Popham, 1987). Chapter Four focuses on the concept of washback and washback potential as it has been investigated in the research literature. However, if the links are strong in an education system, in other words if there is a high degree of alignment, then the three components of an education system will be tightly held together to provide the students with a better opportunity to attain the standards and learn, and

thus, a consistent message to all participants within the education system is delivered (Herman et al., 2007; La Marca, Redfield, Winter, Bailey & Hansche, 2000; Linn, 1994). Therefore, within an ideal education system, what students are assessed on ought to be derived from what is expected of them as specified by the set standards which guide or determine? what they are taught.

Possible questions related to high-stakes/standardised tests that may be addressed in alignment research include: 'Have standardised tests led to changes in teachers' instruction?' (Porter, Smithson, Black, & Zeidner, 2007), and 'Do mandated tests narrow the curriculum?' (Au, 2007). Alignment research also examines: (a) possible assessment or instructional deficiencies; (b) whether the curriculum has been oversimplified (Linn, 2000); and (c) whether students have received a fair chance to learn the material on which they were tested (Winfield, 1993). The results of alignment studies can guide policy-makers, test developers, and educators to make modifications in order that content standards, assessment and instruction are connected in order to support students' learning (Herman et al., 2007; Martone & Sireci, 2009). Furthermore, it can be argued that the process of alignment research itself may guide educators to visualise how assessment can be connected to what happens in classrooms to a greater extent (Herman et al., 2007; Martone & Sireci, 2009).

Moreover, alignment research can also inform the public of how any testing instrument is or is not supporting what is supposed to take place in classrooms and of any possible changes that are necessary for an education system (Herman et al., 2007; Martone & Sireci, 2009). The alignment between standards and assessment can be seen as fundamental for:

1. The success of an education system (Webb, 1997);
2. Students' learning (Anderson, 2002; Biggs, 2003; La Marca, Redfield, & Winter, 2000);

3. Accountability of decisions (Daggett, 2000; Haertel & Herman, 2005; La Marca, 2001);
4. Evaluation of education reform (Herman et al., 2007; Martone & Sireci, 2009; Näsström, 2008); and
5. Validation of test score interpretation (Fox & Cheng, 2007; La Marca, 2001).

Within the measurement literature, alignment between standards and assessment is commonly investigated (Bhola et al., 2003; Herman et al., 2007). Alignment research has also analysed the congruency between standards and teaching as well as between assessment and teaching (Porter, 2002). Acknowledging that tests are expected to cover a representative sample of the content standards rather than all the standards, alignment research has the advantage of illustrating which content standards were covered, and the curricular objectives that have been tested (Martone & Sireci, 2009). Having mentioned that alignment research shares certain goals and evaluation methods with the traditional content validity research (Section 3.1.4), the following sub-section provides an overview of the conceptualizations of validity which have informed the present study. This is followed by a description of the three main evaluation methods/models for measuring alignment. Notably, the three methods have much in common, but on closer inspection each method has its own relative strengths and weaknesses.

### **3.2.2. A Contemporary Definition of Validity**

Validity, in its simplest form, has traditionally been concerned with the question of whether or not a test measures its intended construct. Thus, some have considered validity to be an “inherent attribute or characteristic of a test and assumes that a psychologically real construct or attribute exists in the minds of the test-takers - if something does not exist, it cannot be measured” (Van der Walt & Steyn Jr., 2008, p.192). However, this traditional view of test validity has been challenged by a new conceptualization in which validity is conceptualised not

as a characteristic of testing, rather it is regarded as a property of the interpretation of test scores (Brualdi, 1999; Guerrero, 2000; La Marca, 2001; Messick, 1989, 1996; Van der Walt & Steyn Jr., 2008). In accordance with Messick (1989), I consider validity as a property of the interpretation of test scores, rather than a characteristic of a test.

One conceptualization of validity is derived from the seminal work of Samuel Messick (1989), who altered the view of validity as a property of the test to validity being closely associated with test score interpretation. Messick (1989) argued that “[v]alidity is an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment” (p.13). In addition, Messick (1989) presents a unified model of the contemporary view of validity, in which he represents “construct validity as a central underlying component, with content criterion validity as aspects of construct validity” (p.20). Messick (1989) proposes a matrix (see Table 3.1) consisting of a four-way classification of validity. This matrix is depicted by the source of justification testing that considers both evidence and consequences, and the function or outcome of both the test interpretation and use.

Furthermore, Messick (1989) emphasizes the social dimension of testing and validity. In other words, the importance of the social consequences and impact of tests on test-takers and society. This social perspective of validity is known as ‘consequential validity’. The coined term consequential validity encompasses concepts that range from the “uses of tests, the impacts of testing on test takers and teachers, the examination of results by decision makers, and the potential misuse, abuse, and unintended usage of tests. In other words, consequential validity implies that tests have various influences both within and beyond the classroom” (Pan, 2009, p. 258). In other words, Messick’s *consequential validity* concept refers to “societal implications of

testing that are only one facet of a broader, unified concept of test validity” (Pan, 2009, p. 258), and under which the washback phenomenon came to be considered as one of the fundamental components of any test validation process (Messick, 1996; Van der Walt & Steyn Jr., 2008; Weir, 2005).

Table 3.1:

Messick’s Facets of Validity (Source: Messick, 1989)

|                                |                            | <b>Function of Testing</b>              |  |
|--------------------------------|----------------------------|---|--|
|                                |                            | <b>Test Interpretation</b>              | <b>Test Use</b>  |
| <b>Source of Justification</b> | <b>Evidential Basis</b>    | Construct validity                      | Construct validity + relevance/utility                       |
|                                | <b>Consequential Basis</b> | Construct validity + value implications | Construct validity + relevance/utility + social consequences |

Based on this description it can be argued that validity is an abstract concept, while the process of operationalising it can be described as validation. Van der Walt and Steyn Jr. (2008) consider that the inductive validation process of tests involves two stages. The first stage is the test results and the evidence these present of the test-takers’ abilities, and the second is the empirical validation process of the methods by which the test judgements were arrived at. The latter stage requires the collection of evidence from various sources including construct validity, content validity, criterion validity and consequential validity. Consequential validity considers test consequences, washback, and impact of tests, and the ethics of testing practices (Fives & DiDonato-Barnes, 2013). In a nutshell, the changes in the conceptualization of validity have shifted the focus from establishing different aspects of validity (for example, content vs. construct validity) to establishing several lines of validity evidence that all add to the validation of test score interpretations (La Marca, 2001).

Furthermore, the evaluation of inferences obtained from test scores starts with the evaluation of the test itself. At this stage, researchers ask several questions: Is the content of the test consistent with the test construct? Does the test measure the intended objectives?; and Does the test content align with the intended curriculum?. These questions are primarily concerned with content validity analysis (Guerrero, 2000; Van der Walt & Steyn Jr., 2008; Martone & Sireci, 2009; Ruhio, Berg-Weger, Yiehh, Lee, & Rauch, 2003; Sireci, 1998).

Therefore, alignment research in the above context can be seen as important for both the validation and interpretation of assessment results (La Marca, 2001). It can be further argued that the alignment between standards and assessment can be both a characteristic of the assessment itself and the evidence in the validation of the test score interpretations (Näsström, 2008). Conforming to the AREA, APA, NCME (2014) definition of validity<sup>15</sup> if the degree of alignment of an assessment is measured before its administration, alignment in such case is a characteristic of the assessment and not a matter of validity. However, if one is providing evidence for assessment results' interpretations then the evaluation of the degree of alignment will be a validity concern (La Marca, 2001; Näsström, 2008).

In an ideal scenario, general test specifications are formulated during the development stage of an assessment, and more importantly the articulated test specifications build on the set standards (Näsström, 2008). A well-written specification will result in a record that, if given to a group of similarly trained test developers working in a similar context, will construct a set of test tasks that are similar in content and measurement characteristics (Davidson & Lynch, 2002; Cheng & Fox, 2017). Therefore, test specifications become critical for: documenting the characteristics of a test to guide test construction; enhancing test objectivity; and for generating a

---

<sup>15</sup>“Validity is defined as “the degree to which accumulated evidence and theory support a specific interpretation of test scores for a given use of a test” (APA, AERA, NCME, 2014).

record of evidence drawn together to address the issue of validity (Guba & Lincoln, 1989). Furthermore, Guba and Lincoln (1989) stress that if this record of specification is maintained as it evolves, it becomes a validity narrative. This record may later be presented for peer-review and as a permanent audit trail (Davidson & Lynch, 2002) or history file (Cheng & Fox, 2017) that can be reviewed by multiple stakeholders in a testing context

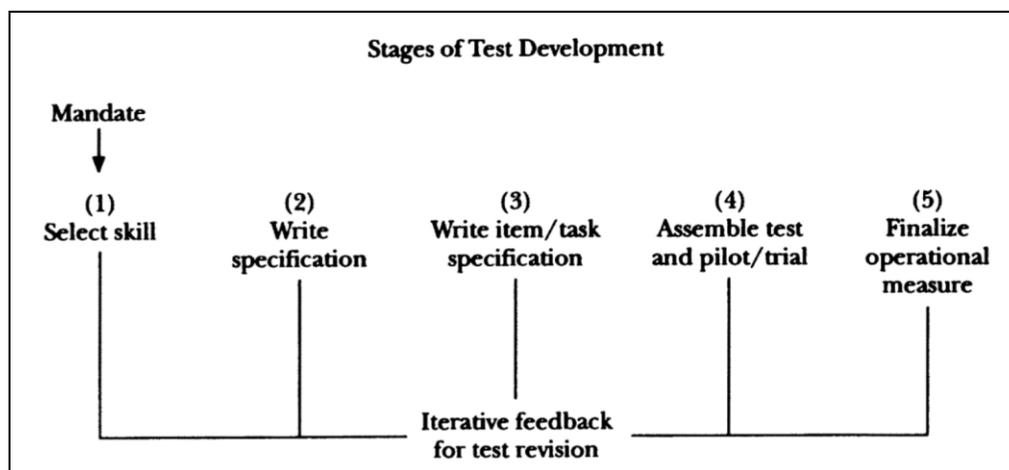


Figure 3.3: The Role of Test Specifications in the Stages of Test Development

Source: Lynch & Davidson (1994, p.729)

It is worth noting that the test specifications for any assessment may change at any point during the test development process, which is an iterative process, working back and forth between the criterion specification and the test tasks (Cheng & Fox, 2017; Davidson & Lynch, 2002). Figure 3.3 illustrates this view. Note that the feedback channel runs under the whole process, and thus, specifications may be influenced and changed by feedback at all stages of test development.

Test specifications within a standard-based educational system centre around the set standards and can build on all the standards or just a sample of the standards. The number of standards incorporated in test specification is primarily determined by the allocated test time, availability of resources, or other restrictions set by the testing body (Näsström, 2008). When

integrating alignment research into a testing context, the evidence provided can possibly differ depending on whether “the assessment specifications include all standards, just a sample of the standards or other knowledge and skills than those in the standards” (Näsström, 2008, p.23). Accordingly, Näsström (2008) argues that assessment specifications and the numbers of standards considered interact to produce four types of different alignment roles in relation to validity. Each case is considered below.

In Case 1, where all the standards are considered in the test specification (see Figure 3.4), test items ought to be a typical sample of all the possible items that assess all the set standards (Näsström, 2008). In a context such as Case 1, where results are generalised and interpretations of students’ attainment of all the standards are considered, alignment can be used as evidence of support. Therefore, alignment research is a validity issue that is similar to content validity (Näsström, 2008).

In Case 2 (see Figure 3.5), where only a sample of the standards is included in the test specification, the test items should be a representative sample of all the likely items assessing only the chosen standards that were articulated in the test specifications. Content validity in the Case 2 scenario, is an “evaluation of how representative the items in the assessment are in relation to the universe of items assessing the standards in the assessment specifications” (Näsström, 2008, pp.25-26). Näsström (2008) notes that alignment in Case 2, still considers all the standards, but it is no longer recognised as content validity. Instead alignment is considered part of construct validity because the test scores are to be generalised to the domain of standards articulated in the assessment specifications and they have to be inferred to devise the chain of inferences on students’ attainment of all standards (Kane, 2006). Alignment in Case 2 can thus be evidence for supporting an extrapolation (Näsström, 2008, p.26).

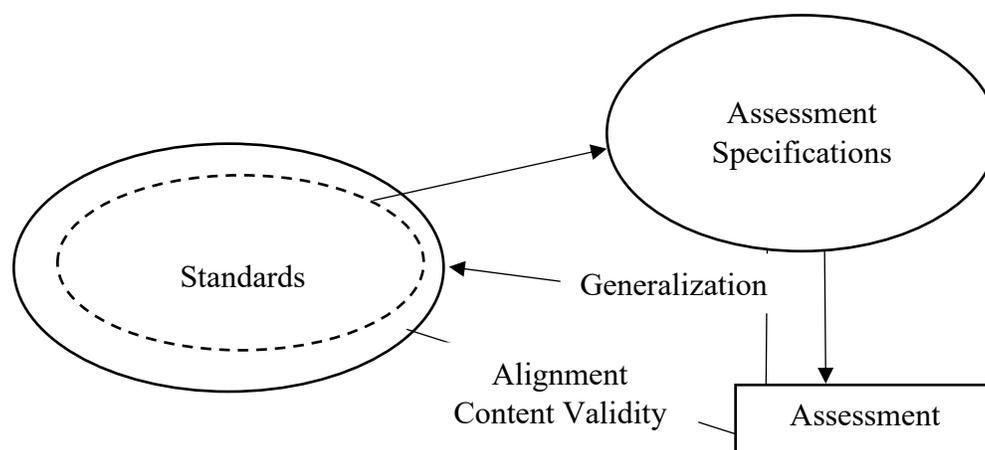


Figure 3.4: Case 1: The Relationship among Standards, Assessment Specifications and Assessment Regarding Validity and Alignment Issues when all Standards are Included in The Assessment Specifications.

Source: Näsström (2008, p. 20)

The Case 3 scenario (see Figure 3.6) includes not only a sample of the standards in the test specifications but also knowledge and skills that are not described and detailed by the standards. Content validity in Case 3 evaluates the test items' degree of representation test specifications, which include the selected standards and all designated knowledge and skills (Näsström, 2008). In addition, alignment within Case 3 is an evaluation of the degree of alignment between assessment and all the standards. Thus, Näsström (2008) argues that content validity and alignment are indicators of two different areas which are not similar. However, alignment can provide “only weak evidence for extrapolations of assessment results to make interpretations about students' attainment of all standards” (Näsström, 2008, p.27). Assessment in Case 3 may influence what and how teachers teach, and in turn alignment can affect consequential validity (Näsström, 2008).

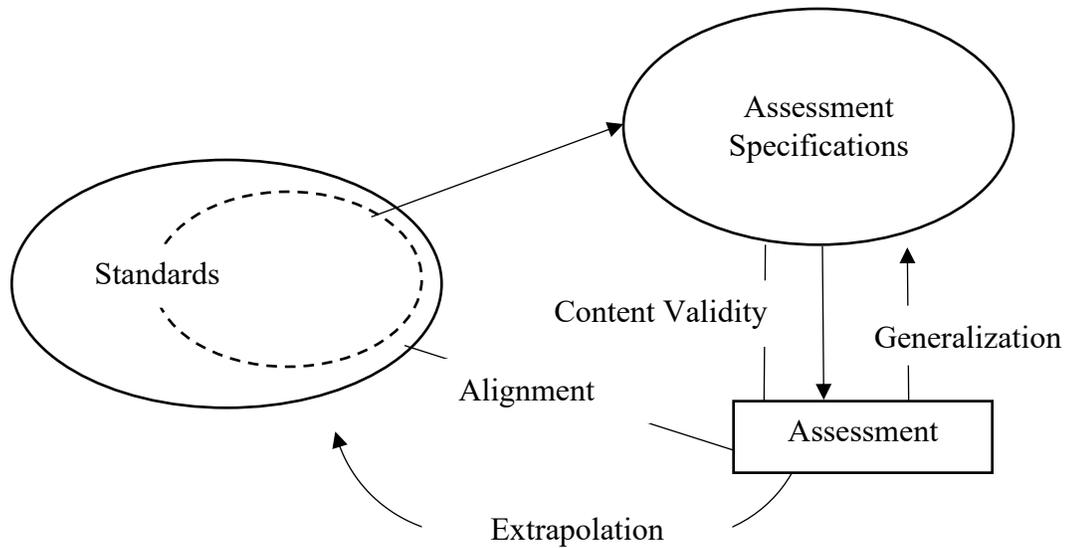


Figure 3.5: Case 2: The Relationship among Standards, Assessment Specifications and Assessment regarding Validity and Alignment Issues when only a Sample of all Standards are Included in the Assessment Specifications.

Source: Näsström (2008, p. 25)

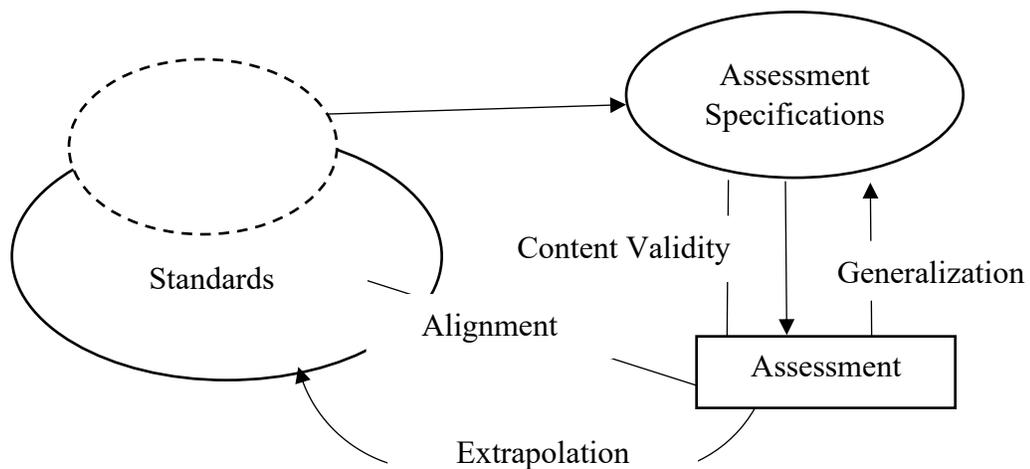


Figure 3.6: Case 3: The Relationship among Standards, Assessment Specifications and Assessment Regarding Validity and Alignment Issues when a sample of all Standards are Included in the Assessment Specifications as well as other Knowledge and Skills.

Source: Näsström (2008, p. 26)

In the last case (Case 4), test specifications do not build on any of the set standards, hence, alignment can “only be used as an indicator of how similar the domain defined by the assessment specifications and the domain of the standards are” ((Näsström, 2008, p.27). It is worth emphasising that alignment can influence the consequences of using an assessment irrespective

of the number of standards included within the test specifications. For example, in cases where there is low degree of alignment between standards and the employed assessment; therefore, student grading, evaluation of reform policies, rewards and sanctions “will have a weak base as regards students’ attainment of all standards” and a “risk that teachers will teach only what is assessed and the students” (Näsström, 2008, p.27). With such consequences it can be noted that alignment affects the consequential validity of any assessment (Näsström, 2008).

In summary, although Webb (1997) notes that “alignment corresponds most closely with content validity and consequential validity” (p.4), the contemporary definition of validity and Näsström’s (2008) four cases encourage researchers to consider and describe alignment as a source of evidence for not only consequential and content validity but also construct validity (Forte, 2016). It is Näsström’s view that guides the research in the present study.

### **3.2.3. Alignment Research Models**

La Marca (2001) states that both the development and application of alignment research resulted from a strong desire to ensure that students’ test scores accurately reflect their performance with regards to expected standards. Thus, systematic procedures for assessing alignment have been established (Ananda, 2003; Bhola, Impara, & Buckendahl, 2003; Olson, 2003; Rothman, Slattery, Vranek, & Resnick, 2002; Webb, 1997, 2002). These systematic procedures for assessing alignment were initially established by Andrew Porter and Norman L. Webb and are widely employed across the United States (Herman, Webb, & Zuniga, 2007).

Bhola et al. (2003) present a detailed overview of the various models to alignment research and classify each model according to the degree of complexity necessitated by each model. The three most popular alignment models are the Webb, Surveys of Enacted Curriculum (SEC) methods; and Achieve models. Alignment in low complexity models is defined as the level of

congruence between test items and the intended content standards. These judgements are arrived at by a panel of SMEs rating the degree of match on a Likert-scale. Such an approach is typically employed by traditional content validity research (Buckendahl, Plake, Impara, & Irwin 2000; Sireci, 1998a, 1998b). Bhola et al. (2003) differentiate between moderate and low complexity models by emphasizing that the former consider test items matches from both the content and cognitive requirements. An example of moderate complexity can be found in the SEC methods in which standards, assessment and instruction are aligned. Finally, the Webb (1999) and Achieve (1998) approaches are examples of high complexity models, as they draw upon additional criteria to provide a broader view of alignment. Notably, recently developed alignment models focus their analysis primarily on judging the degree of alignment between standards and assessment, and only a small number of these models consider the degree of alignment between standards and teaching (Näsström, 2008). Possible reasons for this include:

- Standards are mostly articulated without taking into account teaching, and/or teaching materials (Jongsma, 1993) and hence, different teaching approaches are anticipated.
- Assessments, and in particular large-scale assessments, are expected to assess the standards as the results are commonly used to provide feedback to students and for evaluating educational reforms (Guskey, 2007; Herman, Webb, & Zuniga, 2007).
- Large scale assessments are believed to influence teaching (Agee, 2004; Gerwin & Visone, 2006; Linn, 2006; Luna & Turner, 2001).

Table 3.2

*Overview of Major Alignment Models (Source: Roach et al., 2008, p.162)*

|   | <b>The Webb Model</b>   | <b>SEC</b>  | <b>Achieve</b>  |
|---|---|---|---|
| <b>Components evaluated for Alignment</b> | Assessments standards   | Assessments<br>Standards and Curricular Materials<br>Classroom Instruction  | Assessments (Items and Item Sets)<br>Standards  |
| <b>Raters or Evaluators</b>               | Alignment panel of 6 to 8 educators with subject area expertise   | Individual teacher (Classroom Instruction);<br>Alignment panel of 3 or more content area specialists  | Alignment panel of 3 or more content area specialists   |
| <b>Alignment Evaluation Process</b>       | <ol style="list-style-type: none"> <li>Panel members are trained to recognize and apply four depth-of-knowledge (DOK) levels.</li> <li>Panel reaches consensus on DOK level ratings for objectives from content standards.</li> <li>Panel members then independently rate the DOK level and corresponding objective from standards for each assessment item.</li> </ol> | <ol style="list-style-type: none"> <li>Teachers complete SEC ratings at the end of the year. Survey includes ratings level of coverage for topics and subtopics taught and the level of cognitive demand for tasks in each topical area.</li> <li>Panel members rate the level of coverage for topics and subtopics and cognitive demand of tasks and activities for standards, curricular materials, and assessments.</li> </ol> | <ol style="list-style-type: none"> <li>Expert panels make consensus judgments regarding the quality of the content and performance match between individual test items and their respective standards. Each item is further evaluated regarding the source of its difficulty.</li> <li>Panels then judge whether entire item sets assess the respective standards with a comparable emphasis and range of expectations. Each set of items is further evaluated regarding the grade-level appropriateness for its span of difficulty.</li> </ol> |
| <b>Breadth criteria</b>                   | Categorical<br>Concurrence,<br>Range of<br>Knowledge,<br>Balance of<br>Representation   | Topic and subtopic categories,<br>Emphasis ratings within Topics  | Content Centrality (Items)<br>Range (Item Sets)<br>Balance (Item Sets)  |
| <b>Depth criteria</b>                     | DOK Consistency   | Cognitive demand categories<br>Emphasis ratings within cognitive demand   | Performance Centrality (Items)<br>Source of Challenge (Items)<br>Level of Challenge (Item Sets)   |

Therefore, it can be argued that a high degree of alignment between standards and assessments is a trustworthy indicator that the educational system is functioning with adequacy (Näsström, 2008). Table 3.2 above provides an overview of the main features of the three main alignment models, which are discussed in the following section.

### **3.2.3.1. The Webb Model**

Webb (1997) developed a comprehensive and complex approach for investigating the level of alignment between assessment and standards. The Webb model has been described as a milestone for examining the degree of alignment between assessment, standards and curricula (Li & Sireci, 2004). Although Webb's alignment approach considers five inclusive dimensions—content focus, articulation across grades and ages, equity and fairness, pedagogical implications, and system applicability (see Table 3.3 for their descriptors)—only the content focus dimension has tended to be used in alignment research (Webb, 1997). The term 'standards' in Webb's model refer to a subject's broad content domain and expected skills, and standards are referred to as 'objectives'. Another important feature of Webb's model is the term 'hit' which refers to any item-objective match. In addition, 'expectations' within the Webb perspective can be expressed in the content standards, which "stipulate the knowledge and skills students are expected to acquire and which the assessment is usually based on, or some other forms of framework such as recommended instructional practices" (Li & Sireci, 2004, p.6).

As illustrated in Table 3.3, the content focus dimension within Webb's model involves the analysis of six subcategories to arrive at a level of alignment between assessment and standards. The six categories are: depth of knowledge (DOK); categorical concurrence; range of knowledge; balance of representation; structure of knowledge; and dispositional consonance

(Lane, 2004; Martone & Sireci, 2009). However, only the first four categories have been the focus of analysis in alignment research and thus are discussed in greater detail below.

Table 3.3:

*The Complete Set of Webb's 1997 Alignment Criteria* (Source: Forte, 2016, pp.6-7).

|  |
|--|
| <p><b>I. Content Focus</b></p> <ul style="list-style-type: none"> <li>A. <i>Categorical Concurrence</i>: correspondence between the topics in the standards and the topics by which assessment results are reported.</li> <li>B. <i>Depth of Knowledge Consistency</i>: ratings of most cognitively demanding assessment activity for a topic within the standards as determined by number of ideas integrated, depth of reasoning required, knowledge transferred to new situations, multiple forms of representation employed, and mental effort sustained correspond to the same type of ratings of most cognitively demanding assessment activity for that same topic within the assessment.</li> <li>C. <i>Range of Knowledge Correspondence</i>: standards and assessments cover a comparable span of knowledge within topics and categories.</li> <li>D. <i>Structure of Knowledge Comparability</i>: the relationships among ideas (e.g., no relationship, equivalent forms of the same idea, connection of many ideas within the content area, and connection of ideas within the content area and with applications to other areas) expressed in the standards are the same as those required to perform successfully on the assessments.</li> <li>E. <i>Balance of Representation</i>: the weight by topic or subtopics in the standards corresponds with their weight on the assessments (weight could be determined by the proportion of activities by topic, proportion of average time allocated to do an assessment activity by topic, or according to some other rule).</li> <li>F. <i>Dispositional Consonance</i>: the desired dispositions toward the content area students are to develop as described in the standards are dispositional qualities are observed, monitored, and reported at designated levels within the system.</li> </ul> <p><b>II. Articulation Across Grades and Ages</b></p> <ul style="list-style-type: none"> <li>A. Cognitive soundness determined by best research and understanding: the expressed or implied underlying theory of how students' learning progresses over time that is represented in the standards is reflected across grades in the assessments.</li> <li>B. Cumulative growth in content knowledge during students' schooling: the expressed or implied understanding of how students' knowledge of content will be structured and will mature over time as represented in the standards is reflected across grades in the assessments.</li> </ul> <p><b>III. Equity and Fairness</b></p> <p>Students are afforded a fair and reasonable opportunity to demonstrate the full level of knowledge expected for all students. Assessment practices are such that variation of assessment results are only a variation in the attainment of expectations and free from being influenced by culture, ethnicity, gender, or any other irrelevant factor.</p> <p><b>IV. Pedagogical Implications</b></p> <ul style="list-style-type: none"> <li>A. Engagement of students and effective classroom practices: Instructional practices most likely to have students fully achieve expectations are the same as the instructional practices most likely to have students adequately demonstrate their attainment of these expectations on the assessments.</li> <li>B. Use of technology, materials, and tools: adequate performance on assessments require students to be accomplished in using the full range of technology, materials, and tools as intended by the expectations.</li> </ul> <p><b>V. System Applicability</b></p> <p>The public, teachers, students, and others within the system view expectations and assessments as closely linked, acceptable, attainable, and important.</p> |
|--|

*Depth of knowledge consistency:* involves the degree of alignment between the cognitive demands of both the standards and the test items (Lane, 2004; Martone & Sireci, 2009; Webb, 1999). Webb explains that the DOK consistency indicates alignment if “what is elicited from students on the assessment is as demanding cognitively as what students are expected to know and do as stated in the standards” (2002, p.5). The fundamental criteria for DOK consistency is that 50% of the items for any analysed objective ought to be at or above the expected cognitive level of knowledge (Lane, 2004; Martone & Sireci, 2009; Webb, 1999, 2002). The 50% criteria is based on the assumption that most cut-off scores require students to obtain and achieve this level in order to pass (Lane 2004; Martone & Sireci, 2009; Webb, 1999, 2002). The cognitive levels of knowledge have been classified by the Webb model into four levels: recall; skill and/or concept; strategic thinking; and extended thinking (see Table 3.4). However, these cognitive levels of knowledge can be modified according to each individual study.

*Categorical concurrence:* is the fundamental requirement in any alignment study, regardless of the complexity level (Lane, 2004; Martone & Sireci, 2009; Webb, 1999, 2002). Categorical concurrence is concerned with comparing the similarity between the content standards’ expectations and the actual assessment under investigation. For this category of consistency, Webb (2002) recommends employing criteria of at least six hits measuring a standard for an effective alignment of a test. For example, if a subject matter has six standards, the test would need at least 36 hits for it to achieve categorical concurrence. The logic behind these criteria is that “at least six items would be needed if students were to receive scores on a standard because fewer than six items would not likely result in scores of sufficient reliability” (Martone & Sireci, 2009, p.1338).

Table 3.4

*Webb's General Descriptions for Depth-Of-Knowledge Levels (Source: from Webb, 2002).*

| Level                          | Description   |
|--------------------------------|---|
| Level 1: Recall                | This level includes the recall of information such as a fact, definition, term, or simple procedure, as well as performing a simple algorithm or applying a formula.  |
| Level 2:<br>Skill/Concept      | This level includes the engagement of some mental processing beyond a habitual response. A Level 2 assessment item requires students to make some decisions as to how to approach a problem or activity. Key words that distinguish a Level 2 item or task include “classify,” “organize,” “estimate,” “make observations,” “collect and display data,” and “compare data.”   |
| Level 3: Strategic<br>Thinking | This level includes items that require reasoning, planning, using evidence, and a higher level of thinking than the previous two levels. In most instances, requiring students to explain their thinking is a Level 3 attribute. Students might also be required to make conjectures or determine a solution to a problem with multiple correct answers at this level.  |
| Level 4: Extended<br>Thinking  | This level includes items that require complex reasoning, planning, developing, and thinking most likely over an extended period of time. At Level 4, the cognitive demands of the task should be high, and the work should be very complex. Students should be required to make connections both within and between subject domains. Level 4 activities include designing and conducting experiments, making connections between a finding and related concept, combining and synthesizing ideas into new concepts, and critiquing literary pieces and experimental designs. |

*Range of knowledge:* this consistency category analyses the breadth of standards as compared to the breadth of assessment (Webb, 1999, 2002). It looks into the number of objectives measured by at least one assessment item. All the objectives within the range of knowledge consistency category are assumed to have equal weighting and cover the necessary skills to achieve the standards (Martone & Sireci, 2009, Webb, 1999). Webb (2002) emphasizes that a test is considered to have achieved sufficient alignment if 50% or more of the objectives within a

standard have been measured by one or more test items. The logic behind Webb's (2002) assigned criteria is that students ought to be tested on at least one half of the curricula content.

*Balance of representation*: analysis is "used to indicate the extent to which degree to which items are evenly distributed across objectives" (Webb, 1999, p.9). This category first focuses on the objectives being assessed by the test items and considers the percentage of objectives being measured in comparison to the number of tested items (Roach et al., 2005).

For objectivity purposes, the four Webb criteria described are numerically rated, quantified, calculated and reported (Case, Jorgensen & Zucker, 2004). In summary, analysing the four categories functions is the means by which researchers arrive at the degree of match between an assessment and the standards within in an education programme.

Webb (1997b) went on to further classify his model of alignment analyses into two categories known as horizontal and vertical. As illustrated in Figure 3.7, horizontal alignment involves the degree of alignment between components at the same level of an educational system, and vertical involves alignment between components at different levels.

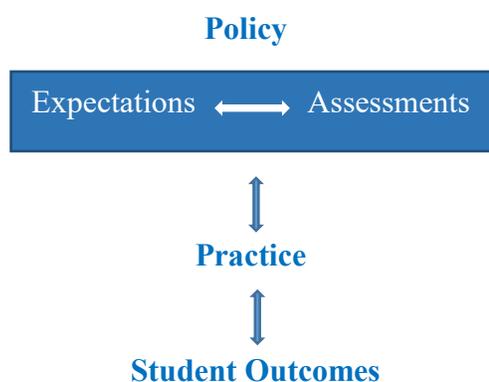


Figure 3.7: Horizontal and Vertical Alignment within an Education System  
Source: Adapted from Webb (1997b, p.1)

Studies that employ the Webb model, generally recruit and train a panel of educators and assessment and curriculum experts. The review panel members are trained to use an analytic and heuristics process to examine the degree of alignment between the target academic standards and assessments (Roach et al., 2008). After the training session is complete, the panel members have the following tasks to complete during the alignment process:

1. Agree on a DOK level for each objective in the target content standards.
2. Rate the DOK level for every assessment task/item.
3. Assign each listed test item with the best corresponding objective(s) from the content standards.

Webb's approach combines both qualitative expert judgment and quantified coding to evaluate the alignment of standards and assessments (Flowers, Browder, & Ahlgrim-Delzell, 2006). Webb's comprehensive conceptualization of alignment has resulted in the model being used by many studies in different contexts, such as Elliott et al. (2001), Roach et al. (2005, 2008), Webb, (2002), and Webb et al. (2002), and its reliability has been well established. Indeed, Webb's model of alignment has been used to determine the alignment of a large-scale testing with the academic standards of more than 20 US states across various subject matters including language, arts, mathematics, social studies and science (Roach et al., 2005; Roach et al., 2008). The results have provided fundamental information for policy-makers that have been later used to: change assessments; modify standards; and validate the degree to which the two educational components were directed towards mutual expectations for learning (Roach et al., 2005). A particular advantage of the Webb model is the presence of specific levels for each of the four criteria in order to help educators and policy-makers decide on the adequacy of the alignment between content standards and assessments. This information can also serve for

determining the subsequent steps necessary for revising the assessment and/or standards (Martone & Sireci, 2009). On these lines, Bhola et al. (2003) and Martone and Sireci (2009) argue that the Webb model is a comprehensive approach for analyzing alignment and it provides a point of reference for SEC and Achieve models.

### **3.2.3.2. SEC model**

The SEC model was developed by Porter and Smithson (2001, 2002). The main features of the SEC model are: a collective language framework for the content of curriculum, instruction and assessment (CIA); alignment statistics; and graphical output of the CIA content (Roach, Niebling, & Kurz, 2008). In addition, strategies for evaluating the degree of emphasis on a number of content topics in all curriculum, instruction and assessment are considered. The matrix employed in this model consists of two dimensions: content topic and category of cognitive demand. In the alignment process, four content experts (primarily teachers) are asked to code each content standard and assessment item into the two-dimensional matrix (Bhola et al., 2003). The elicited data from the coding matrix are then converted into graphs, charts and the analysis of the data results in statistics and an alignment index (Case et al., 2004; Porter, 2002). Case et al. (2004) further note that the SEC model can be adapted to evaluate other pillars of an education system; for example, classroom instruction.

### **3.2.3.3. Achieve model**

This model was developed in 1998 by the University of Pittsburgh (Roach et al., 2008). The model employs an “assessment-to-standards alignment protocol to provide a qualitative as well as quantitative analysis of alignment featuring judgments about the quality and rigor of individual test items and sets of items” (Roach et al., 2008, p.167). The panel members are expert reviewers consisting of classroom teachers, curriculum developers, and SMEs. Their

expert judgements come into play when they rate the degree of content centrality, performance centrality balance, and range (Roach et al., 2008). Overall, and in accordance with Resnick et al. (2004), the alignment protocol employed within this model asks the panel experts to gather information regarding a test's breadth in respect to the set standards, sampling of content and skills, and level of difficulty. The application of the Achieve model results in a thorough and informative report detailing each component of the alignment process. The report even highlights areas of suggestions on how to improve both assessment and standards (Martone & Sireci, 2009). The Achieve model has been used by 14 states in the USA to examine the overall degree of alignment between state assessment and set standards (Roach et al., 2008). As noted by Resnick et al. (2004), this model was designed to address the following three questions: (a) does assessment measure only content and skills reflected in the standards? (b) does the assessment fairly and effectively sample the important knowledge and skills in the standards? and (c) is the assessment sufficiently challenging? The assessment items for standards are evaluated against five criteria that cover content centrality, performance centrality, challenge, balance and range. *Content centrality* focuses on the match between assessment items and the content objectives (Forte, 2016, Greive, 2012). Like Webb's DOK consistency level, *performance centrality* criterion considers the concurrence of assessment items' cognitive complexity and the assigned objectives. The *challenge* criterion examines to what degree an assessment item has a "range of difficulty that is both matched to the level of difficulty in the objective and appropriate for the target students" (Greive, 2012, p.19). *Balance* considers to what extent the assessment items cover the same range of skills and knowledge in the content domain. Finally, *range* looks at whether the assessment items test the same range of knowledge and skills as defined in the content standards (Forte, 2016, Greive, 2012). Table 3.5 provides further information on the

three alignment approaches in relation to their key features, and the time required for analysis and training.

Table 3.5

*Summary of the Three Alignment Models (Source: Case et al., 2004, p.9).*

| <b>Model</b> | <b>Key Features</b>   | <b>Review and Analysis Time</b>  | <b>Training Time</b>                   |
|--------------|---|--|--|
| Webb         | <ol style="list-style-type: none"> <li>1. Qualitative ratings</li> <li>2. Quantitative results</li> <li>3. Can measure inter-reliability and variation in alignment statistics</li> </ol>                                 | <p>1 day per team of knowledge (Multiple grades); 1-month turnaround for analysis and report. Using the Webb Alignment Tool, alignment analysis is a two-part process. In the first part, reviewers reach consensus on the DOK levels for the objectives under the standards. This takes about 2 hours. In part 2, reviewers code a test to the standards by identifying the DOK for each item and the corresponding objective. This takes from 60 to 90 minutes. [Complete reports are produced in one to two weeks].</p> | ½ day to train reviewers               |
| SEC          | <ol style="list-style-type: none"> <li>1. Content matrix</li> <li>2. Measure of alignment highly predictive of student achievement scores</li> <li>3. Information can be applied to help educators and schools</li> </ol> | <p>1 day per team for coding items and benchmarks in matrix; ½ day for readers to complete survey on instruction; 1 week for analysis and report.</p>  | ½ day to train reviewers               |
| Achieve      | <ol style="list-style-type: none"> <li>1. Reviewers need to make inferences</li> <li>2. In-depth review</li> <li>3. Provides technical reports</li> </ol>   | <p>Alignment review takes 1 day per test; report and analysis takes 1-1½ months.</p>   | Has a pool of highly trained reviewers |

### **3.3.3.4. Commonalities and differences across the three approaches of alignment**

Although different alignment models have been developed they share a number of commonalities. Each approach to alignment has a different aim with different strengths and limitations. From the summary highlighted in Table 3.5, and in accordance with Bhola et al. (2003), Case et al (2004), Martone and Sireci (2009), and Näsström (2008), it can be argued that:

1. In the three models, the assessment items are categorised by at least two criteria and are related to the standards.
2. The most commonly used criteria in alignment models are content and cognitive complexity. However, the definition of these two criteria vary across the models.
3. The Webb approach offers the most thorough quantitative results and offers clear guidelines about the adequate levels of alignment.
4. Through employing Webb's four categorical analysis the weak and strong dimensions of alignment can be established.
5. The Achieve model builds on the foundations of the Webb model and delves further into the source and level of challenge dimensions to capture the quality of test items, which overcomes a limitation of the Webb model.
6. The Achieve model has the advantage of providing qualitative as well as quantitative information about the alignment and the value of the matches.
7. A clear cut-off for each criterion is found within the Achieve model, as alignment is considered from a holistic perspective rather than an analytical perspective (Resnick et al., 2004).
8. The SEC model is the only alignment approach that takes into account the instructional component of an education system. The approach provides a simple comparison of the

three pillars of an education system within a nation or across states and districts.

However, unlike the Webb and Achieve models, the SEC model does not provide much detail about the quality of the alignment.

The three models presented are the most advanced and comprehensive models in relation to alignment research. Each model has been successfully employed in different education contexts (Case et al., 2004). In choosing a model, one has to consider the financial, time and personal resources at hand, as well as the main aim of the research and the importance of the information that will be provided by the results.

Overall, the systematic study of alignment can be time consuming and costly. However, alignment research is empowering as it provides information for revising both assessment and content standards, and thus, the benefits outweigh the costs (La Marca, 2001).

In this chapter, I have discussed literature relating to the alignment between standards, instructional activities and materials, and assessment, which constitute the three fundamental components of an education system (Näsström, 2008). Alignment, in the view of the current study, resides in the triangular relationship between the following three components: standards, testing (e.g., high-stakes tests – rSECEE), and curriculum (.i.e., the standardised textbook within the Libyan context).

With very little or no information known about the degree of alignment between the rSECEE and Libya's EFL content standards, Phase I aimed to answer the first research question that guided this study: **To what degree is the rSECEE aligned with Libya's EFL content standards?**

By answering this question, I would be able to indicate whether the rSECEE adequately measures the concepts and skill areas represented in the Libyan EFL content standards, and if the

rSECEE captures the meaningful aspects of student thinking and learning. Plus, I will be able to assess whether standards, testing (e.g., high-stakes tests – rSECEE), and curriculum are all directed towards mutual expectations for students' learning (Martone & Sireci, 2009; Roach et al., 2005). Therefore, the empirical evidence derived from this study may provide useful information for policy-makers that can be used (if needed) to change assessment procedures or alter standards.

In order to measure the degree of alignment between standards and testing for the Libyan education system, the Webb (1997) model was employed, because as noted earlier its additional criteria provides a much broader view of alignment. In addition, Webb's alignment model meets Green's (2007) requirements that washback studies undertake a detailed analysis of the target testing instrument and a thorough evaluation of its alignment with the set standards. In addition, the inclusiveness of the Webb model allows it to be adapted to other educational contexts such as the Libyan context in which alignment studies are needed (Impara, 2001).

Given the dominant role high-stakes testing plays within the Libyan education context, high-stakes testing, and the literature on washback and high stakes testing is reviewed in the following chapter.

## Chapter IV

### Washback and High-Stakes Testing

Because the present study seeks to understand the relationship between the degree of alignment among the components of an educational system (i.e., standards, curriculum, and testing) and the washback of a high-stakes test on the classroom, the chapter considers the literature on washback and high-stakes testing. The definition of washback and a brief overview of how the washback phenomena evolved is provided. This is then followed by washback trends and a short description of what I consider to be influential models of washback.

#### 4.1. Washback

##### 4.1.1. Definition of Washback

As discussed in Chapter One in the fields of applied linguistics and language testing definitions of washback, sometimes referred to as ‘backwash’ in general education (Biggs, 1996; Hughes, 1989, 2003), are almost as numerous as the researchers who write about it. These definitions range from the simple to the very complex. However, the most common labels attached to the concept of test influence are ‘washback’ and ‘impact’. The term ‘washback’ is rarely found in dictionaries, but derives from the literal meaning of the word ‘backwash’; the backwash flow or movement of something onto something else, such as the backwash of sea water onto a beach (Saville, 2009). Thus, washback in the assessment and language testing literature is figuratively used to denote the ‘backward flow’ of effects from testing onto teaching and learning.

Messick (1996) describes washback as the “extent to which tests influence language teachers and learners to do things they would not otherwise necessarily do” (p.241). It is worth noting that

Messick (1996) situates the two concepts of ‘washback’ and ‘impact’ within one theoretical concept of consequential validity (see Section 3.2.2), which refers to the importance of the social consequences and impact of tests on test-takers and society. Using consequential validity, the washback phenomenon came to be considered as one of the fundamental components of any test validation process (Messick, 1996; Van der Walt & Steyn Jr., 2008; Weir, 2005). Messick (1996) further warns that it cannot just be claimed that washback exists because particular features exist in the test’s operational context. In his words, “it is problematic to claim evidence of test washback if a logical or evidential link cannot be forged between the teaching and learning outcomes and the test properties thought to influence them” (Messick, 1996, p.247).

Other researchers including Alderson and Wall (1993), Bailey (1996, 1999), Cheng and Curtis (2004), and Hughes (1989), offer a simpler definition of washback as the influence of testing on teaching and learning. For example, according to Wall (1997) washback refers to “the effects of tests on teaching and learning” (p.291). However, Shohamy (1993, 2001) regards washback as an intentional exercise of power over educational institutions with the goal of controlling the learners’ and teachers’ behaviour.

In comparison, Bachman and Palmer (1996) include washback within their concept of ‘test impact’, which is a term that refers to the effects that tests have on individuals (teachers and students), on education systems, and on society at large. Similarly, Shohamy (2001) differentiates washback from impact by using impact as an umbrella term under which washback falls. She explains that washback can occur at micro-levels of an educational context involving participants such as teachers and learners, but impact can occur at the institutional and social level, i.e. the macro level of a context.

In contrast, Hamp-Lyons (1998) and McNamara (1996, 2000) define the two terms, ‘washback’ and ‘impact’, differently. According to McNamara (1996), impact refers to the effects of language tests on the macro-levels of education and society, while washback is the effect of tests on the micro-levels of language testing and learning, i.e., inside the classroom.

Thus, certain researchers refer to ‘impact’ and ‘washback’ as unique but related concepts, while others refer to one as a subset of the other, and still others use the terms distinctly. Therefore, the different definitions of washback in the language testing literature reveal differences in scope. In the context of the current study, and because of the nature of this research, the term ‘washback’ fundamentally refers to “[t]he influence of testing on teaching and learning” (Bailey, 1996, p.259). Thus, the focus is on the washback at the micro-level of rSECEE on Libyan Grade 12 teachers and students. In accordance with Spratt (2005), using washback in this context allows for “both the accidental and the intentional effects of washback” (p.8), while the door is left open on “whether washback is positive or negative” (p.8), and which factors that may bring about which effects. However, I use the term ‘washback’ or ‘backwash’ in its original form when quoting other authors, while ‘consequences’ is used in a similar manner to Cheng et al. (2015) who refer to it as “the direct results of (mis)use of test scores” (p.438). For example, if a Grade 12 Libyan student fails the second sitting of an rSECEE, a consequence would be retention (i.e. holding the student back one year and repeating Grade 12). Finally, the term ‘impact’ hereafter is used without any technical connotations and is only used as a synonym for effect.

#### **4.1.2. Washback: Positive, Negative, Neutral or Both**

Washback often implies the movement in a certain direction (Green, 2007, 2013; Spratte, 2005; Wall, 2012). The movement of washback in a particular direction can describe the

relationship between testing and teaching. As noted by Pearson (1988), “public examinations influence the attitudes, behaviour, and motivation of teachers, learners and parents, and, because examinations come at the end of a course, this influence is seen as working in a backwards direction: hence the term ‘washback’” (p.98; also, see Figure 4.1).

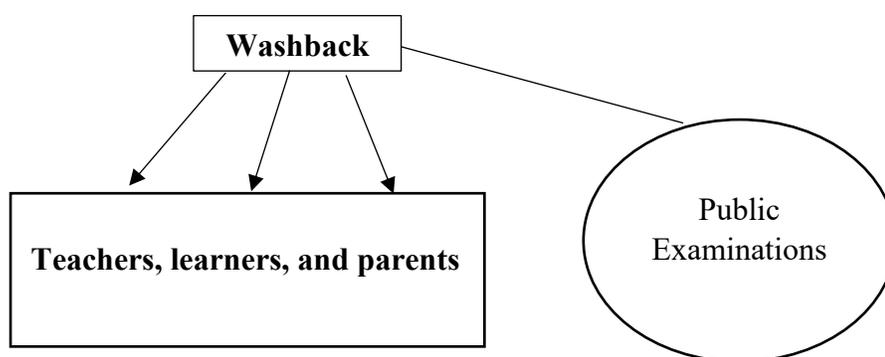


Figure 4.1: The Washback Effect of Public Examinations.  
Source: Pearson (1988, p.98)

In addition, washback has also been perceived of as *bipolar*; either positive or negative (see Green, 2007). The washback effects of a test would be considered positive if a test brought about the desired changes. However, a washback effect would be considered negative if learning principles were not reflected in the test as expected (Pearson, 1989), or the test brought about undesirable effects on teaching and learning (Alderson & Wall, 1993). Therefore, washback is often evaluated as positive or negative according to whether it encourages or discourages forms of teaching or learning judged to be appropriate (Green, 2007, 2013).

Research in the language testing field has presented evidence similar to that documented in the general education literature<sup>16</sup>, which argues that language examinations may bring about negative washback on both teaching and learning.

<sup>16</sup> The literature review provided in this Chapter explores the international literature on standardised testing and its impacts, and is dominated by research from the United States, the United Kingdom and China, because these

#### 4.1.2.1. Negative Washback on Teachers and Teaching

The concern of educational researchers about the negative washback of high-stakes testing on teachers and teaching dates back to the late 20<sup>th</sup> century. Madaus (1985) argues:

Frankly, I am alarmed that policy makers, when they mandate tests for decisions about graduation, promotion, or merit pay increments, are attempting to legislate as the engine or primary motivating power in the educational process. My concerns are rooted in history, which provides many examples of the results of such a policy: The exam become the master and not the servant of the educational process; it invariably leads to cramming; it narrows the curriculum, it constraints the creativity and spontaneity of teachers and students; and final it demeans the professional judgement of teachers (p.6).

Proponents of high-stakes testing programmes are not comfortable with the idea of publishing the limitations of implementing such testing systems in education systems because they have “become powerful tools for educational reform, tools that leverage money, sanctions, curricula, and public esteem for schools” (Paris, 2000, p.6). Many large-scale high-stakes tests are implemented on the assumption that this testing practice improves instructional practice. In other words, the assumption is that “testing drives much of what teachers do, and so curricular and instructional change will occur if and when the tests change” (Grant, 2000, p.2). Based on this assumption, policy-makers believe that there is “potential for big pedagogical changes with a modicum of effort: change the test and one changes teachers’ practices” (Grant, 2000, p.2).

Whether or not these testing programmes bring improvements in the what and the way teachers teach, and teachers’ perceptions about the change in instruction practice has been the focus of

---

nations have the longest histories of standardised high-stakes testing (see, for example, Polesel, Dulfer & Turnbull, 2012; or Cheng, 2008).

many researchers, including Cheng (1997, 2004, 2005) Cimbricz (2002), Heubert and Hauser (1999), Grant (2000), Green (2007), Shohamy et al (1996), and Urdan and Paris (1994).

Teachers' perceptions about the high-stakes testing practices reported within the international literature (such as Fairbairn & Fox, 2009; Linn, 2000; Urdan & Paris, 1994) include yielding inadequate results for parents and the public, being unfair to minority students such as second language speakers (Fox & Cheng, 2007), and unworthy of the time or cost that is spent on them.

Teachers have also reported that because of the stakes associated with standardized testing practices and the pressure of producing high test scores they have a "diminished sense of professional worth and feeling of disempowerment and alienation" (Blazer, 2011, p.6).

Likewise, in Taylor et al.'s (2002) study, where teachers' perspectives on high-stakes testing in Colorado were examined, 81% of the teachers reported that they had experienced a decrease in faculty morale, which they ascribed to the existence of a high-stakes testing programme. High-stakes testing often sends very strong signals to schools about what they should be teaching and what students ought to be learning, and in response, schools tend to teach what is being assessed (Herman, 2004; Koretz, Mitchell, Barron, & Keith, 1996; Stecher, Barron, Chun, & Ross, 2000).

It is reported that some teachers have (and may still) directly assist students during the administration of high-stakes tests, or in extreme cases answer questions on their behalf (Jacob & Levitt, 2003; West, 2007). A possible rationale for the negative reports on high-stakes testing impact may be the stakes involved. Paris (2000) highlights that there are high-stakes for teachers and school administrators because students' scores are reported in media and can be used by education officials as a basis for rewarding high achieving schools and placing sanctions on lower achieving schools. Studies cited in the education literature have repeatedly confirmed that "increasing the stakes attached to tests can change what is taught and how it is taught and

adversely affect the quality of classroom practice” (Blazer, 2011, p.2). The pressures that teachers experience when teaching within a high-stakes testing context forces the teacher to choose between teaching the prescribed curriculum, or preparing students for tests that can involve different curricula content and skills. In addition, multiple research studies have reported that many teachers who are working within a high-stakes testing context, narrow and distort the scope of the curriculum (Au, 2007; Debray, Parson, & Avila, 2003; Luna & Turner, 2001; Madaus, 1988; Popham, 1987; Shepard, 2002; Smith, 1991a, 1991b), and sacrifice innovative teaching methods and resort to traditional teaching approaches such as teacher-centered lecturing in order to *cover the curriculum* (Brimijoin, 2005; Luna & Turner, 2001; Prodromou, 1995).

The narrowing of the curriculum within a large-scale high-stakes testing context can occur in four different ways:

1. *Exclusion of non-tested subject areas*: research documents that time spent on non-tested subject matter areas is either reduced or eliminated (Amrein & Berliner, 2002; Cimbricz, 2002; Shepard, 2002; Smith, 1991a, 1991b).
2. *Exclusion of non-tested topics within subject areas*: studies record that teachers teaching towards a large-scale high-stakes test tend to focus pedagogical attention on the curriculum components that are to be tested and other untested content is simply marginalised (Cimbricz, 2002; Minarechová, 2012; Stecher, 2002);
3. *Acclimating teaching style to the testing format*: the stakes of large-scale testing programmes have encouraged teachers to use instructional practices and materials that mirror both testing format and content (Cimbricz, 2002; Rottenberg & Smith, 1990; Stecher, 2002). Gayler et al. (2003) state that “[m]any educators report that exit tests and high-stakes testing are squeezing out any content not covered by the tests, encouraging

breadth of coverage instead of depth, and promoting a curriculum sequence and a pace that are not appropriate for some students” (p.10). It is also noted that instead of building a thorough understanding of concepts and principles, teachers tend to focus more on the memorisation of the target content (Green, 2013; Roach, Niebling, & Kurz, 2008; Smith 1991b; Smith & Rottenberg, 1991).

4. *Test preparation practice*: teaching time that is devoted to test preparation practices such as doing mock and former test papers is commonly known in the literature as *teaching to the test* (Madaus, 1988; Minarechová, 2012). Research, including Cimbricz (2002) and Stecher (2002), reports that teachers working within a large-scale high-stakes testing context tend to engage in test preparation activities that “promote ways of teaching that are often boring and neglectful of problems and issues concerned with race, class, gender, and sexuality” (Grant, 2004, p.7). Although engaging in test preparation practice and familiarising students with testing procedure and content may increase test scores, it adds no-to-very-little improvement for students in meeting the set expectations (Blazer, 2011). Madaus et al. (2009) document similar findings, and further add that teachers have even taken advantage of recess time to prepare for the tests.

Other documented examples of negative washback include the existence of a learning atmosphere that is full of “high anxiety and fear of test results” for teachers and their students (Shohamy et al., 1996, p.309, see also Alderson & Wall, 1993; Cheng, 1997, 1998, 2004, 2005).

The seminal empirical research on washback was conducted by Wall and Alderson (1993). They undertook a two-year longitudinal study on the consequences of implementing a revised O-level examination on teaching methods in Sri Lanka. Along with modified textbooks,

the revised version of the test was introduced in order to promote communicative English language teaching. The revised examination placed greater emphasis on speaking, reading, and writing skills. Although through observation Wall and Alderson (1993) found that the language learning activities and the design of classroom tests were influenced by the revised test, they found no difference in the way the teachers taught over the two years of the study. The English classes continued to employ a grammar and teacher-centered approach.

After Wall and Alderson's (1993) seminal work, a number of empirically-based washback research projects reported that there were negative washback effects on some teachers but not all (Green, 2007; Watanabe, 1996, 2004; Yu 2010). It was established that the degree and kinds of washback occur through "the agency of various intervening bodies and are shaped by them. An important and influential agent in this process is the teacher" (Spratte, 2005 p.26), suggesting that teachers "face a set of pedagogic and ethical decisions about what and how best to teach and facilitate learning" (Spratte, 2005 p.26).

#### **4.1.2.2. Washback on Learners and Learning**

From the research cited in the education literature, it can be argued that high-stakes/standardised testing programmes do very little in improving students' knowledge and skills (Marchant, 2004). Without the act of constructive feedback, the testing process is merely an activity whereby students demonstrate their knowledge and skills rather than their learning (Marchant, 2004). In line with Marchant (2004) students may use high-stakes test results as a labelling indicator of whether they are "smarter", or necessarily "dumber" (Marchant, 2004, p.2). Students' fear of high-stakes testing and its consequences is probably due to what is at stake (Paris, 2000). Due to the stakes of standardized tests, some students have come to undervalue

both schooling and learning and the mere focus of schooling may become “whether this will be tested or not” (Paris, 2000). Extensive research on the matter of school retention establishes that it can be a faulty policy that brings with it damaging and disturbing life-long effects (Anderson, Whipple, & Jimerson, 2002; Paris, 2000). Furthermore, Futrell and Rotberg (2002) show that there is a notable increase in students’ dropout rate as a by-product of high-stakes testing programmes.

It can be further argued that any narrowing of the curriculum and providing day-to-day activities that closely resemble the test format may deny students learning opportunities in many ways. First, learning isolated facts and skills can be difficult, because without context there is no meaningful way to accumulate or systematize information and make it easy to remember (Shepard, 1991). Second, learning “decontextualized skills means that the subsequent application of skills to real world problem becomes a separate and difficult learning hurdle” (Shepard, 1991, p.233). Third, students may be mistakenly categorised as high-achievers not because of their actual competence, but rather for being engaged in extensive test preparation activities (Volante, 2004). Last, learning may be undervalued, and thus, discourage students, especially those who are most in need of improvement (Fairbairn & Fox, 2009; West, 2007).

Moreover, the impact of large-scale high-stakes testing continue to be far reaching for students. As early as the 1990s, the education literature documents that high-stakes testing can negatively affect students’ health and well-being (see, for example, (Polesel et al., 2012). International research provides evidence that high-stakes testing can affect students’ mental and physical state by:

- Lowering self-esteem (Marchant, 2004; Perrone, 1991);

- Causing considerable amount of stress, anxiety, and sense of uselessness (Gregory & Clarke, 2003; Lewis, 2000; Stiggins, 1999);
- Undermining self-worth (Gregory & Clarke, 2003); and
- Viewing test performance as a prediction of future failure and hardship (Reay & Wiliam, 1999).

In essence, criticisms with regards to mandated large-scale high-stakes testing have mainly focused on the potential of reducing the “opportunity to learn”, and the actual reliability of the tests in meeting the set aims and their effects on teachers and teaching, learners and learning. Skrtic (1995) describes the testing system as a “machine bureaucracy” (p. 199), that regulates and controls schools, teachers and students with little to no consideration of the value of teaching and learning. Plus, as the stakes become more “consequential for students”, the pressure to do well increases correspondingly (Paris, 2000, p.3). It is the negative consequences that impact students that are most concerning, and those consequences can be substantial (Marchant, 2004).

#### **4.1.2.3. Positive Washback on Teachers and Learners**

Positive washback can apply not only to teachers but also their students. These characteristics include students having a more positive perspective towards learning, and therefore a greater willingness to work harder, and teachers fulfilling their teaching goals, as well as, adhering to their students’ needs and interests (Mahmoudi, 2014). Researchers, such as Andrews, Fullilove and Wang (2002), Bachman and Palmer (1996), Bailey (1996), Hughes (1989) and Hsu (2009), believe that it is possible to bring about positive washback through changing an examination. Suggestions for promoting positive washback include: incorporating testing abilities that one wishes to encourage; ensuring that teachers and students are knowledgeable about the test

construct and objectives (Hughes, 1989); and involving teachers and test-takers in the design and development of the test (Bachman & Palmer, 1996).

Similarly, within the educational literature advocates of high-stakes testing programmes (e.g., Braun, 2004; Phelps, 2006; Vogler, 2003) also argue that a comprehensive testing programme can bring many benefits to an educational context. Testing specialists, government officials, policy-makers and some researchers promote high-stakes testing programs (such as standardized tests) as “scientific and valid assessments of academic achievement” (Paris, 2000, p.5). For example, Gradwell (2006), Fickel (2006), van Hover (2006) and Wolf, Wolf, and Carpenter (2002), argue that high-stakes testing has little-to-no negative effects on what teachers do in the classroom. Braun (2004) and Williamson, Bondy, Langley, and Mayne (2005) note that high-stakes testing may lead to an advantageous learning experience and potentially positive educational outcomes. An example of the positive impact of high-stakes testing on the curriculum at the classroom level is illustrated by Vogler (2003). He found that social studies teachers added social studies content to their curriculum in response to the high-stakes test which primarily tested writing ability rather than social studies content knowledge. Moreover, Paris (2000, pp.6-7), who presents an opposing view, highlights the claimed benefits of high-stakes testing as:

1. Students will work harder and learn more when they have high-stakes tests.
2. Students are motivated to do their best and score well in high-stakes test.
3. Doing well in high-stakes test leads to feelings of success and doing poorly leads to increased effort for learning.
4. Students and teachers need high-stakes tests to know what is important to learn and teach.
5. Teacher need accountability through high-stakes test to motivate their teaching.

6. High-stakes tests are good measures of the curricula that students are taught in schools.
7. Tests are a “level playing field” and provide an equal opportunity for all students to demonstrate their knowledge.
8. High-stakes are good measures of an individual’s performance and are affected little by differences in students’ motivation, interest, emotionality, language, and background.
9. Parents understand high stakes and to interpret their children’s scores.

Phelps (2006) adds that “[High-stakes testing] information can be used for diagnosis (of individual students or teachers, of schools, and of school programs)” (p.19), and can help maintain the set academic standards and identify students at risk (i.e. student who are not meeting the standards), in order to offer them the support that fits their individual needs.

Moreover, standardised testing programmes are said to be more reliable than individual teacher grading and evaluation, which may consider non-cognitive factors such as class participation, perceived efforts, and progress over the course period (Phelps, 2006). On similar lines, other researchers including Sloane and Kelly (2003) and Amrein and Berlinern (2002) argue that testing can play a role in improving student motivation and providing students with adequate information about their knowledge and skills.

The following section provides a brief overview of how the washback phenomena evolved and highlights what studies to date reveal about the complexity of the washback phenomenon. Before proceeding, it is worth noting that due to the somewhat short history of empirical research on washback in applied linguistics and language testing, and in accordance with Cheng’s (2014) distinction, this study considers research published before and during the 1990s as ‘previous’ and that published in the 2000s as ‘current’ research. This type of distinction is very important in the field of applied linguistics and language testing because it was only in the 1990s that washback

came to be recognised as an important, wide-ranging phenomenon (Cheng, 2014). Thereafter, in the 2000s there was a significant increase in the volume of empirical research from different parts of the world (Cheng, 2008, 2014).

#### **4.2. Overview of Washback**

Until the early 1990s washback was believed to relate to the test design itself, in that “if it is a good examination, it will have a useful effect on teaching; if bad it will have a damaging effect on teaching” (Heaton, 1990, p.16). In terms of its evolution, Gosa (2004) describes washback as a concept that has passed through three phases of development. She describes the pre-1990s era as ‘the myth phase’ and defines it as the period when writers claimed the existence of test influence without any empirical evidence. The next phase was described as ‘the metaphor phase’, which started in 1993 with the publication of Alderson and Wall’s seminal paper. Alderson and Wall (1993) were the first to question the extant descriptions of the nature of test influence. Their work offered a re-conceptualization of the washback phenomena and they concluded that there is not always a linear relationship between the design of tests and what takes place in the classroom in terms of teaching and learning (Tasagari, 2011). This “marked a significant development in shaping the construct of washback studies for the field of language testing” (Cheng, 2014, p.4) in the early and mid-1990s. To systematise the investigation of washback, Alderson and Wall (1993) proposed 15 hypotheses that involve the differences between the effects of examinations on attitudes and on the content of teaching and learning, and between the impacts on methods and the impacts of processes (Green, 2013). After 1993, Gosa (2004) believes research on test influence entered ‘the reality stage’. This era was characterised by its evidence-based research and an attempt to develop models that could help explain the nature of washback.

Following Alderson and Wall's (1993) call for an increase in empirical studies, a number of empirical-based washback research projects investigated a range of tests in different parts of the world (such as Andrews, 1995; Andrews, Fullilove, & Wong, 2002; Cheng 1997, 1998, 2004, 2005; Fox & Cheng, 2007; Roberts, 2000; Saif, 2006). For a non-exhaustive review on previous and current washback studies see the summary table in Appendix C.

### **4.3. The Washback Trends in Language Testing Literature**

Following Morrow's (1986) claim that testing researchers need to go into classrooms "in order to observe the effects of their tests in action" (p.6) and Alderson and Wall's (1993) call for an increase in empirical studies, a number of empirical-based washback research projects investigated a range of tests and their influence on teachers, methodology and teaching material (Cheng, 2008, 2014; Green, 2007). By reviewing both previous and current washback studies, Cheng (2014) argues it was evident that research priorities that characterised the previous studies during the 1990s focused primarily on classroom-based participants (teachers and learners). Evidently, teachers are the most visible stakeholders to be affected by test washback as they are "the 'front-line' conduits for the washback process related to instruction" (Bailey, 1999b, p.17). The importance of teachers and teaching in washback processes was strongly emphasized by Alderson and Wall's (1993) in a number of their washback hypotheses (1, 3, 4, 7, 9 and 11). It can be argued that during the 1990s teachers and their teaching were (and still are) the most frequently studied of all the stake-holders involved in the washback process. Along these lines, Shohamy (1992) emphasizes the central role of teachers when identifying the possible conditions that can lead to negative washback:

After all, when reliance is on tests to create change; when emphasis is mostly on proficiency and less on the means that lead to it (i.e., what takes place in the classroom as

part of the learning process); when tests are introduced as authoritative tools, are judgmental, prescriptive, and dictated from above; when the writing of tests does not involve those who are expected to carry out the change - the teachers; and when the information tests provide is not detailed and specific and does not contain meaningful feedback and diagnosis that can be used for repair, it is difficult to expect that tests will lead to meaningful improvement in learning (p.514).

In one of the previous washback studies, Lam (1994) examines 61 teachers' perceptions of changes brought about by the New Use of English, and, more specifically, the influence it had on how teachers taught and prepared students for the public examination. From his results, Lam (1994) concludes that it is not sufficient to change exams: "[t]he challenge is to change the teaching culture, to open teachers' eyes to the possibilities of exploiting the exam to achieve positive and worthwhile educational goals" (p.96).

It is evident from the review of washback studies that although language learners are the participants whose lives are mostly directly influenced by language testing, the research during the 1990s did not focus on learners and their learning processes. Cheng's (1998) was probably the first washback study during the 1990s to have examined the washback of Hong Kong Certificate of Education Examinations in English on learning, as well as learners' perceptions and practices. Cheng (1998) found that even though learners changed their learning beliefs after the content of the test had been modified, they reported maintaining their initial learning processes, learning strategies, and individual motives to learn English. Furthermore, at that time there was very limited research that examined the affect of testing on other stakeholders in the wider context, such as parents, exam developers, administrators, and so on (Andrews & Yu, 2011; Baily, 1999; Cheng, 2008, 2014; Saville, 2009).

However, since the turn of the century, researchers have expanded their washback perspective and looked at issues of context in order to capture the complexity of the phenomenon (Cheng, 2005; Gu, 2005; Pan, 2010; Qi, 2003, 2004; Shih, 2006; Wall, 2005). Current washback studies have taken a broader view by including stakeholders at the macro-level. For example, Saville and Hawkey's (2004) in their study of the University of Cambridge Local Examinations Syndicate highlight the wide range of stakeholders at the macro-level (Booth, 2012). Saville and Hawkey's (2004) list of stakeholders operating in a testing community is very similar to the one proposed by Rea-Dicken (1997), and this list includes test-takers, teachers, parents, test administrators, test users, teacher educators, sponsors and funding bodies, government bodies, the public, various national and international examination authorities, members of working parties and curriculum committees,

Thus, current washback research not only focuses on the micro-level of test influence, but also on the different levels of stakeholders within an educational context, as can be seen in Table 4.1. The Table summarises some current washback research, detailing the area of research, researchers involved, and the geographical context in which the research took place. Table 4.1, although not exhaustive, illustrates how since the turn of the century washback research continued to examine test consequences on teachers and teaching. However, there is a notable increase in the number of washback studies that examine learners and their learning and the recognition that the roles of various stakeholders within an educational context are now considered important. These include textbook publishers, parents, administrators, test designers, government official, and inspectors. The research of Cheng (2005), Gu (2005), and Wall (2005) is of particular significance because they investigate the impact of high stakes public

examinations on a huge number of stakeholders (at both the micro- and macro-levels) within the educational contexts of Hong Kong, China, and Sri Lanka.

Overall, it can be argued that since the turn of the century the washback phenomenon has been examined much more seriously, both theoretically and empirically. The shift in washback research trends, from primarily focussing on teachers and teaching to considering different levels of stake-holders within an educational context, can be attributed to the change in the conceptualization of washback. The current conceptualization of washback now considers test influence to exist not only in classrooms, but that its ramifications occur in the wider social and political environment (McNamara & Roever, 2006).

Table 4.1

*A Selective Review of The Current Washback Trends*

| <b>Researcher(s)</b>              | <b>Research Area</b>  | <b>Research Context</b> |
|-----------------------------------|---|-------------------------|
| Andrews, Fullilove, & Wang (2002) | Learners and learning outcomes  | Hong Kong               |
| Ferman (2004)                     | Teachers, learners, learning, and inspectors  | Israel                  |
| Borrows (2004)                    | Teachers and teaching   | Australia               |
| Cheng (2005)                      | Micro- and macro-level stakeholders including teachers, learners, and textbook        | Hong Kong               |
| Cheng, Andrews & Yu               | Parents and learners  | Hong Kong               |
| Fox & Cheng (2007)                | Learners  | Canada                  |
| Gosa (2004)                       | Learners' test preparation  | Romania                 |
| Green (2007)                      | Teachers, learners and test design  | United Kingdom          |
| Gu (2005)                         | 4,500 stakeholders including administrators, teachers, and students.                  | China                   |
| Han, Dai, & Yang (2004)           | 1,194 teachers  | China                   |
| Hawkey (2006)                     | Textbook publishers   | United Kingdom          |
| Hayes & Read (2004)               | Teachers and learners   | China                   |
| Leung, & Andrews (2012).          | Textbooks   | Hong Kong               |
| Nazari (2005)                     | Learners' beliefs   | Iran                    |
| Pan (2010)                        | Multiple stakeholder businesses, government, administrators, educators, and students. | Taiwan                  |
| Qi (2003, 2004, 2005, 2007)       | English inspectors, teachers, learners and test developers                            | China                   |
| Qian (2014)                       | Teachers' professional development  | Hong Kong               |
| Read & Hayes (2003)               | Textbooks   | New Zealand             |
| Scott (2007)                      | Teachers, learners and parents  | United Kingdom          |
| Shih (2006)                       | Multiple stakeholders including department heads, teachers, learners, spouses...etc.  | Taiwan                  |
| Tsagari (2007)                    | Textbooks   | Greece                  |
| Wall (2005)                       | Micro and macro level stakeholders including teachers and learners                    | Sir Lanka               |
| Xie & Andrews (2012)              | Learners test preparation   | China                   |
| Yu (2005)                         | Textbooks   | China                   |
| Mizutani (2009)                   | Teachers and learners   | New Zealand             |

#### 4.4. Washback Models

A number of researchers have developed models that aim to conceptualize washback more precisely, and to inform research agendas (Saville, 2009). The publications of Alderson and Hamp-Lyons (1996), Bailey (1996, 1999), Burrows (2004) Cheng (1997, 1998, 1999, 2000, 2005), Hughes (1993), Green (2007) and Watanabe (1996, 2004) are of particular significance. Although a full description of these models is beyond the scope of this study, it is worth discussing the contributions of Hughes (1993), Bailey (1996), Alderson and Hamp-Lyons (1996), Watanabe (2004), Burrows (2004), Cheng (2005) and Green (2007), in chronological order.

##### **Hughes (1993)**

Hughes (1993) proposed an influential process model of washback. Hughes added to Alderson and Wall's (1993) research by illustrating possible mechanisms through which washback can occur. In his model, he highlighted three major areas that can be affected by the nature of a test in an education context: participants; processes; and products. According to Hughes (1993), participants are classroom teachers and students, education administrators, textbook developers and publishers. Process refers to "any action taken by the participants which may contribute to the process of learning" (Hughes, 1993, p.2), such as material development, syllabus design, changes in teaching methodology, and the use of test-taking strategies. Finally, product was described by Hughes (1993) as "what is learned and the quality of the learning" (p.2). According to this model, the nature of the test may directly influence the perception and attitudes of its participants towards teaching and learning. Accordingly, these perceptions and attitudes may then affect what the participants do; actions which may contribute to or impede the learning outcomes.

### Bailey (1996)

Combining Alderson and Wall's (1993) washback hypotheses, with Hughes's (1993) distinction between participants, process, and product, Bailey in 1996 proposed a new model of washback (see Figure 4.2).

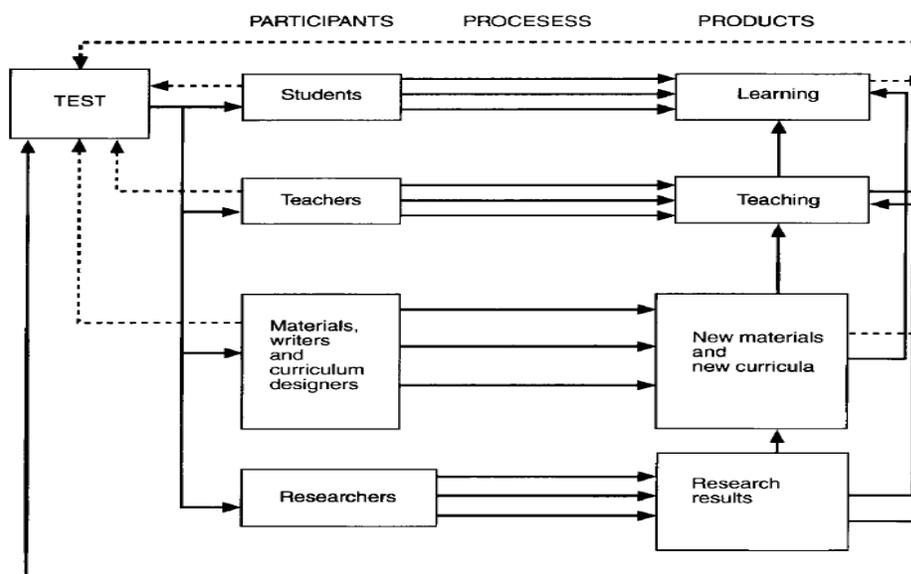


Figure 4.2: Basic Model of Washback

Source: Bailey (1996, p.264)

The model represents the direct influence of tests on various participants, who engage in a number of processes that may result in products specific to different participants. The dotted lines in figure 4.2 represent the possible influence that participants and products may have on the test itself. Furthermore, Bailey (1996) differentiates between 'washback to the learners' and 'washback to the program'. Linking with Alderson and Wall's hypotheses (in particular 2, 5, 6, 8, and 10), she describes test influence on test-takers as 'washback to learners', while 'washback to the program' is related to Alderson and Wall's hypotheses (in particular 1, 3, 4, 8, 7, 9, and 11) and describes test influence on teachers, administrators, curriculum developers, and counsellors. It can be argued that Baily's (1996) concepts of 'learner washback' and 'program washback' overlap to a degree with Bachman and Palmer's (1996) micro- and macro-levels of

washback, although the latter includes the influences of individual teachers under the micro-category.

### **Alderson and Hamp-Lyons (1996)**

Around the same time as Bailey (1996), Alderson and Hamp-Lyons (1996) contributed to the conceptualization of washback by arguing that washback varies across individual teachers and learners. Alderson and Hamp-Lyons's (1996) study was influential in the 1990s because it was the first to investigate the washback of TOFEL (Wall, 2012). The authors collected data based on interviews with teachers and learners, as well as classroom observations in order to challenge the common perception that TOFEL examinations tended to have a negative influence on language teaching classrooms. They compared TOFEL preparation classes with non-TOFEL classes and found that TOFEL influenced what and how teachers taught in the classroom, but the degree and type of TOEFL effects differed amongst teachers. Their research concludes that simple patterns of washback hypotheses are "too naïve" (Alderson and Hamp-Lyons, 1996, p.280), and that test influence on classrooms is complex. Furthermore, they argue "the existence of the test by itself does not guarantee washback, either positive or negative" (Alderson and Hamp-Lyons, 1996, p. 281), and where it does exist it will vary from one person to another. Finally, Alderson and Hamp-Lyons (1996, p. 296) further argue that the amount and type of washback will vary according to a range of factors that include:

- the status of the test (the level of stakes);
- the extent to which the test runs counter to current practice;
- the extent to which teachers and textbook writers think about the appropriate methods for test preparation; and
- the extent to which teachers and textbook writers are willing and able to innovate.

**Watanabe (2004)**

Watanabe (2004) conceptualizes washback in terms of five dimensions: specificity, intensity, length, intentionality, and value. Specificity refers to the type of effect. A general washback effect could be produced by any test, whereas a specific washback effect may “relate to only one specific aspect of a test or one specific test type” (Watanabe, 2004, p.20). Intensity describes whether washback has a strong or a weak effect. If it is a strong effect then the test determines everything that happens in the classroom. However, if a test has a weak effect it will determine only part of the classroom activities (Watanabe, 2004). Length refers to the duration of any washback effects, intentionality simply refers to whether washback is intended or unintended, while value indicates whether washback is positive or negative (Watanabe, 2004).

In addition, Watanabe (2004) argues that features of teaching or learning related to test design, status, and stakeholders may be affected by tests and factors which mediate the process of washback. Watanabe (2004) also aimed to isolate the features of tests that can be responsible for what took place in classrooms. He stressed that washback would possibly exist if:

- teaching, learning, and/or textbooks are different in exam-preparation and non-exam preparation classes taught by the same teacher; and
- teaching, learning, and/or textbooks are similar in exam-preparation classes that are taught by different teachers.

**Burrows (2004)**

Acknowledging the fundamental relationship between assessment and curriculum (see Section 3.1), Burrows (2004) uses curriculum innovation in her examination of the washback effect of classroom-based assessment on the Australian adult Migrant English Program. In washback research, Wall (2005) pioneered the employment of innovation theory in relation to

curricular change in order to help explain how teachers respond in different ways to revised or new tests (Green, 2013). Innovation in the domain of foreign language teaching programmes is defined by De Lano, Riley and Crookes (1994) as:

an informed change in an underlying philosophy of language teaching/learning, brought about by direct experience, research findings, or other means, resulting in an adaptation of pedagogic practices such that instruction is better able to promote language learning as it has come to be understood. (p.489)

Regularly in the literature on language teaching, innovation is concerned with how to bring about pedagogical change. In this context, innovation may be viewed as a vehicle through which teachers can increase their awareness, become better acquainted with the profession, augment their knowledge, and improve their practice in the field (De Lano et al., 1994). According to the literature, there are several factors that can bring about educational change. These include: critical incidents, which, according to Griffin (2003), are situations that allow people to become aware of their implicit beliefs, research, and change agents (De Lano et al., 1994). In terms of change agents, both Atkin (1992) and Whitehead (1989) argue that teachers attempt to make changes once they recognise the presence of a gap or inconsistency between their objectives, principles, and their current practices. Furthermore, it is posited by Hashweh (2003) that teachers experience changes constructively when they hold positive qualities such as a motivation to learn and passion. Similar to Hashweh (2003), Menges (1997) emphasizes the importance of teacher motivation in promoting educational change. Another factor that may trigger change in teachers is *reflection*. Through reflection teachers may become conscious of and critically consider their implicit ideas and practices, and, thus, may construct alternative knowledge, beliefs, and practices (Postholm, 2008).

Burrows (2004) employed Lambright and Flynn's (1980) concepts that define the social roles which participants of a social system assume in relation to other stake-holders. Lambright and Flynn (1980, p.251) proposed that stakeholders can relate to one another as “adopters, implementers, resisters, clients, suppliers, and entrepreneurs”, i.e., change agents in Markee's (1997) terminology. Adopters are participants of a social system who adopt an innovation. Implementers are participants, such as teachers, who make innovations work at the classroom level, whereas, resisters are those who oppose an innovation (Markee, 1997). Importantly, when teachers “implement curricular guidelines at the level of syllabus design, they also manage change in their own classroom – that is, they also act as change agents” (Markee, 1997, p.4).

Burrows (2004) also employed the notion of “models of teacher assessment” (McCallum, Gipps, McAlister & Brown, 1995, p.61) to analyse and interpret her observational data. Teachers' response to a new national assessment programme in England and Wales were examined by McCallum et al. (1995). The findings of their study point to three models of “practice which varied along the dimensions of systematicity, integration with teaching and ideological underpinning” (Gipps, McCallum & Brown, 1996, p.168). The three models were termed *Intuitives*, *Evidence Gatherers* and *Systematic Planners*. Teachers who were minimal adopters of national assessment procedures were identified as *Intuitives*. *Intuitives* resist the innovative assessment practice because it is viewed as a “disruption to intuitive ways of working” (Gipps et al., 1996, p.168) which necessarily implies “radical change in behaviour” for them. McCallum et al. (1995) argue that teachers' resistance may be partly derived from their view of teaching and the prevailing *culture of assessment* (McCallum et al., 1995). *Evidence Gatherers*, however, are those adapt the requirements of an innovative implemented assessment practice. In other words, they are “rational adapters” (McCallum et al., 1995, p.67), who are not

prepared to let an assessment implementation to be “the be-all and end-all” (McCallum et al., 1995, p.71). Teachers can be rational adapters because they value the importance of teaching over the assessment, and that students “learn what is taught and only what is taught; thus, assessment follows teaching in order to check that the process is going to plan” (Gipps et al., 1996, p.169). Finally, Systematic Planners are those who implement the set assessment practices which, in turn, informs their task decisions and classroom practice (McCallum et al., 1995).

Burrows’ (2004) findings led her to conclude that washback is a complex and contested matter and how teachers react to tests is highly individual, and, in accordance with Wall (1996), washback is a form of educational change. Her results provide evidence for the notion that there is a “degree of choice involved in washback” (Burrows, 2004, p.125). This is because, as argued by Burrows (2004), if it is possible resist the effect of a curricular implementation upon classroom practices, then it is possible “to choose whether the implementation of an assessment system or test will have a washback effect” (Burrows, 2004, p.125).

As a result of these findings, Burrows (2004) revised the ‘black box’ washback model, which considers the existence of “individual responses” to washback and that “single washback response is not inevitable” (p.126). Burrow’s proposed washback model (see Figure 4.3) considers both curriculum innovation and teachers’ BAK (Woods, 1995), and assumes washback as a type of educational change that considers teachers’ responses in relation to the educational behavioural models proposed by Lambright and Flynn (1980), Markee (1997), and McCallum et al. (1995).

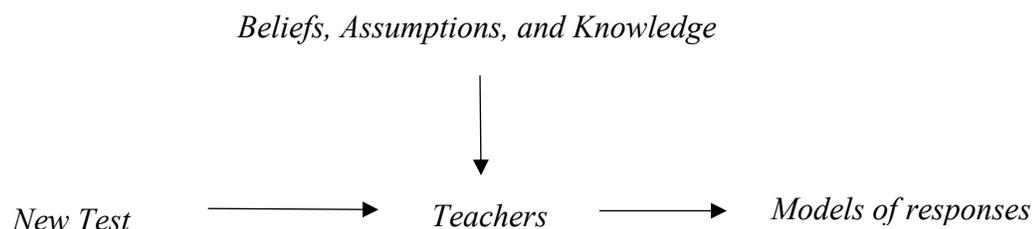


Figure 4.3: Proposed View of Washback: A Curriculum Innovation Model.

Source: Burrows (2004, p.126)

### **Cheng (2005)**

In accordance with the previous discussion, Cheng (2004) confirms the “complex nature of washback effects” (p.162), and that washback is an “educational phenomenon in which change is central” (p.164). Furthermore, in her 2005 longitudinal study, Cheng collected both attitudinal and observational data from schools, and linked behaviours of both teachers and learners to the wider reform objectives of education authorities. Her study concludes that what tended to change in language classrooms was the content rather than the teaching methodology. Hence, Cheng’s (2005) study adds further to the re-conceptualization of the washback phenomena.

### **Green (2007)**

Green’s (2007) influential study investigated how test preparation practices improved students’ IELTS writing scores. In the study, Green (2007) developed a comprehensive model to account for the nature of washback, which Saville (2009) argues extends Baily’s (1996) model by focusing on the target test in itself. Green’s (2007, 2014) model sets out how washback occurs through direction, variability and intensity (see Figure 4.4.). Simply put, Green’s model starts with the test design characteristics (the focal construct) and then measures the extent of correspondence (hence alignment) between the test design and the skills developed in the

curriculum or the target language use domain. It is argued that a well-designed instructional programme considers the relationship between what learners are expected to learn (i.e. the focal construct) and the content of assessment (Green, 2013, 2014). As Saville (2009) argues, the test design is related with the washback direction which has beneficial or damaging effects on both teaching and learning.

Importantly, Green's model considers washback as not just simply related to test design, but also relates test design issues to contexts of test use that can be "realised through and limited by the participant characteristics" (Green, 2007, p.25). The differences of the washback participants' (including material writers, teachers, students, and course administrators) perceptions regarding test demands, importance and difficulty, and their willingness to embrace its importance (hence variability) "will moderate the strength of any effect, [hence, intensity], and, perhaps, the evaluation of its direction" (Green, 2007, p.25). This approach moves the model of washback from "'a recipe' for achieving positive washback, towards a descriptive and partially explanatory tool addressing what goes on in order to cause the various washback effects" (Saville, 2009, p.32).

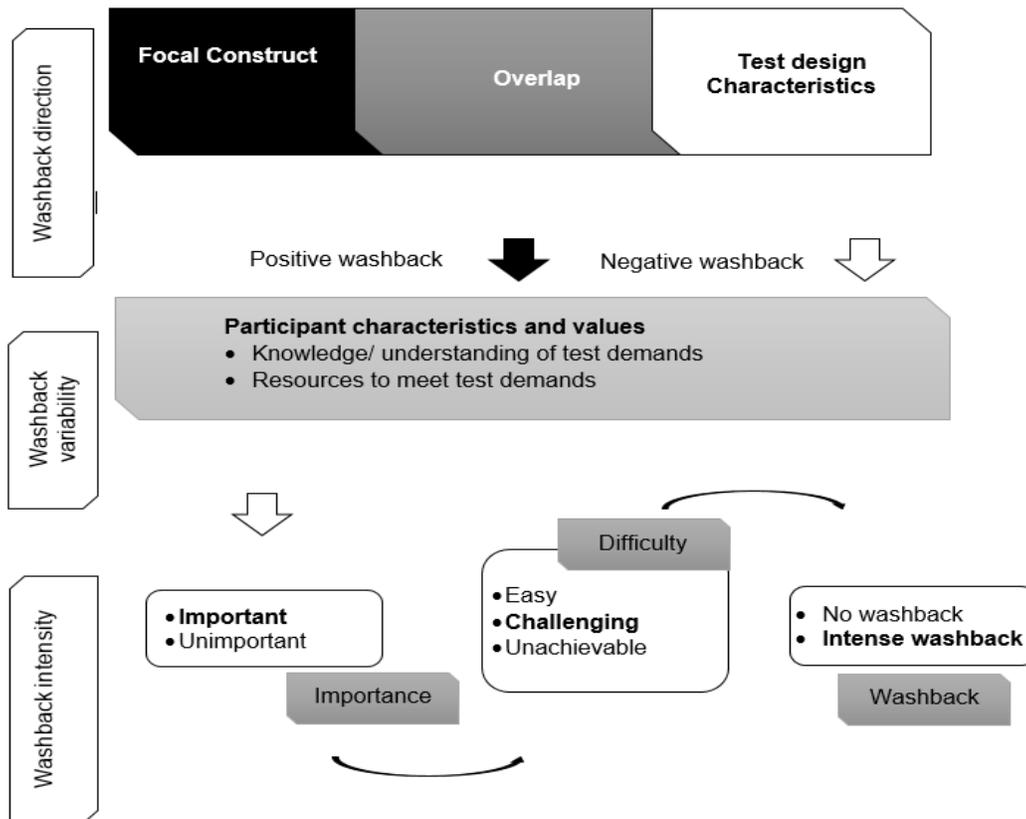


Figure 4.4: Washback Causes and Effects (Source Green 2014, p. 87 which was based on Cambridge English Language Assessment (2007) IELTS Washback in Context: Preparation for Academic Writing in Higher Education, p.24.)

## Summary

In short, since Alderson and Wall's (1993) paper the conceptualization of washback has continued to evolve. Thus:

- 1) Washback is no longer considered a simple concept; instead it is a broad, multi-faceted and complex phenomenon that can vary in both form and intensity (Cheng, 2005).
- 2) Factors in the education system that contribute to the complexity and unpredictability of washback include the status of the language in an educational system, the purpose of the test, the test format and the tested skills, the status of the target language tested (Shohamy

et al., 1996), resources (Chapman & Snyder, 2000), instructional material (Wall & Horak, 2007), and commercial test-preparation materials (Cheng, 2005).

- 3) There are factors that impede the operation of the intended washback, and there is a limited relationship between the test content and the teaching method employed within the language classroom. These factors include: a teacher's beliefs, assumptions and knowledge (BAK, Wood, 1996); a teacher's previous education and academic background; teaching style; teaching experience; and the type of teaching methodology that teachers employ (Alderson & Hamp-Lyons, 1996; Cheng, 1999, 1998; Green, 2007, 2013, 2014; Spratt, 2005; Wall & Anderson, 1993; Watanabe, 1996; see also Appendix D for a summary of all the factors identified by empirical studies as affecting the degree and kinds of washback). Basically, "tests have impact on *what* teachers teach but not how they teach" (Wall & Anderson, 1993, p.68).
- 4) Washback research has established the effects of washback on the classroom including, the curriculum (Alderson & Wall, 1993; Cheng, 1997; Read & Hayes, 2003); materials (Andrews, 1994; Andrews, Fullilove & Wong, 2002; Nikolov, 1999; Read & Hayes, 2003; Watanabe, 1996); teaching methods; feelings and attitudes; and learning (Cheng, 2004, 2005; Cheng et al 2015; Cheng and Fox, 2017; Fox & Cheng, 2007, 2016; Spratte, 2005).

From the above discussion and in accordance with Alderson in his forward to Cheng and Watanabe (2004), washback research nowadays no longer asks "does washback exist?" but rather "what does washback look like? And what brings washback about?" (Cheng & Watanabe, 2004, p. ix). The consistency of the findings in the literature documenting the negative washback of high-stakes testing programmes on teachers and teaching, and learners and learning, raises

profound concerns regarding the Libyan context. Therefore, it is important to investigate the extent to which these results can be generalised to the Libyan context and the recently rSECEE.

Given the important decisions resting on the results of the rSECEE, it is important to understand how well the test is performing within the Libyan context, and to judge if decisions based on the rSECEE is resulting in best practice. Therefore, the Phase II of the current study investigated the second research question: **What is the nature and scope (Cheng, 2004) of the washback (if any) of the rSECEE on the Libyan EFL Grade 12 classroom?**

As discussed above, there are factors (including a teacher's beliefs, assumptions and knowledge (BAK), a teacher's previous education and academic background; teaching style; teaching experience; and his/her students' perceptions (Spratt, 2005, p.29) that can influence intended washback and its intensity. In the Libyan context, it is necessary to examine teacher beliefs because teachers are not blank-slate individuals who implement curriculum policies and objectives handed down by their administrators. Instead, they are entities who filter, absorb, and employ the curriculum policy based on their own beliefs and educational context (Freeman & Richards, 1996; Woods, 1996).

Both the Hughes (1993) and Green (2007) models inform the current study's collection of teachers' and learners' accounts of the rSECEE. These accounts reflect their views of the high-stakes test, and as argued by Hughes (1993) and Green (2007), participants' views can influence the washback direction. This in turn can help me determine whether the washback of the rSECEE is either negative or positive. Furthermore, Burrows (2004) and the curricular literature suggest the degree of teacher variability. Teachers are unique and individual, and are the primary mediators of the curriculum. This is of particular importance given the findings of the present study (see Chapters Seven and Eight).

The following chapter presents an overarching review of the methodology which includes the research questions being addressed in this study, research design overview, a brief description of the research instruments and data collection procedures and analysis.

## **Chapter V**

### **Methodology**

In order to explore the relationship between the degree of alignment among the components of an educational system (i.e., standards, curriculum, and testing) and the washback of a high-stakes test on the classroom, a mixed-methods research design was employed. The method is discussed in this chapter, which first presents the research questions and the presuppositions (Sections 5.1 and 5.2). An overall description of the research design (Section 5.3), and the specifics of the methods used in both phases of the study are then presented. These include a description of the participants, sampling methods, data collection procedures, and data analysis procedures (Sections 5.5, 5.6, 5.7, and 5.8). Reference to how the reliability and trustworthiness of the research was maintained are presented in Section 5.9.

Importantly, as this is a complex multi-phased mixed methods study which requires detailed description of the employed methods, I opted to provide the reader with an overarching review of the methodology in this chapter, and a more detailed description of the data collection and analysis procedures for Phases I and II in Chapters Six and Seven respectively. This decision was taken on the basis that it may help reduce the reader's cognitive load and allow him/her to process the information more effectively.

#### **5.1. Research Questions**

The study asked the following four research questions:

- 1. What is the relationship between the degree of alignment and the washback of the rSECEE? What are the implications of this relationship for key stake-holders (e.g. policy- makers, test developers, teachers, and students)?**

This question guided the researcher to explore:

- 2) **To what degree is the rSECEE aligned with Libya's EFL content standards?**
- 3) **What is the nature and scope (Cheng, 2004) of the washback (if any) of the rSECEE on the Libyan EFL Grade 12 teaching and learning?**

The third research question was addressed through the following sub-questions:

- d) What evidence is there of washback of the rSECEE on teachers i.e., how does the rSECEE influence teachers' accounts of teaching and testing (i.e., external and internal testing)?
- e) To what extent does the rSECEE appear to influence teachers' teaching practices?
- f) How does the rSECEE influence learners' accounts of learning?

In accordance with Cheng (2004) teachers' knowledge and understanding are operationally explored through their accounts of aspects of classroom teaching in relation to the rSECEE.

Aspects of teaching that were the focus of investigation included teachers' accounts of:

1. The reasons behind the change in examination format and content;
2. The test format;
3. Any necessary extra work or pressure in teaching towards rSECEE;
4. Any changes in teaching methodology employed; and
5. Any challenges whilst teaching.

Plus, Students' knowledge and understanding are operationally explored through their accounts of the underlying principles of the rSECEE, and students' views about the rSECEE.

Learning practices are operationally defined as what students say they do in order to prepare themselves for the rSECEE and to improve their level of English.

## 5.2. Presuppositions of the rSECEE Operation

### Phase I

In this phase, it is presupposed that:

There is a weak alignment between the Libyan EFL standards and the rSECEE which may result in negative washback.

### Phase II

Based on Onaiba's (2014) findings and the review of the literature, it is presumed that the Grade 12 EFL classroom is characterised by:

1. Teaching activities that indicate negative washback;
2. Extensive referencing to the examination content and format;
3. The grammar-translation method being the prevailing method of instruction;
4. Increased teacher-centered pedagogy to cover the breadth of test-required information and procedures;
5. Increased attention on grammar and vocabulary learning;
6. Little or no focus on the listening, writing, or speaking skills;
7. A reduction in the amount of instruction and a narrowing of the curriculum;
8. An increase in test format and content practices;
9. The learners pay a great deal of attention to grammar and vocabulary learning;
10. Learners may lack oral communicative competence;
11. Learners are not active in class and do not assume responsibility for their own learning;
12. Learners extensively practice outside the classroom past SECEE papers.

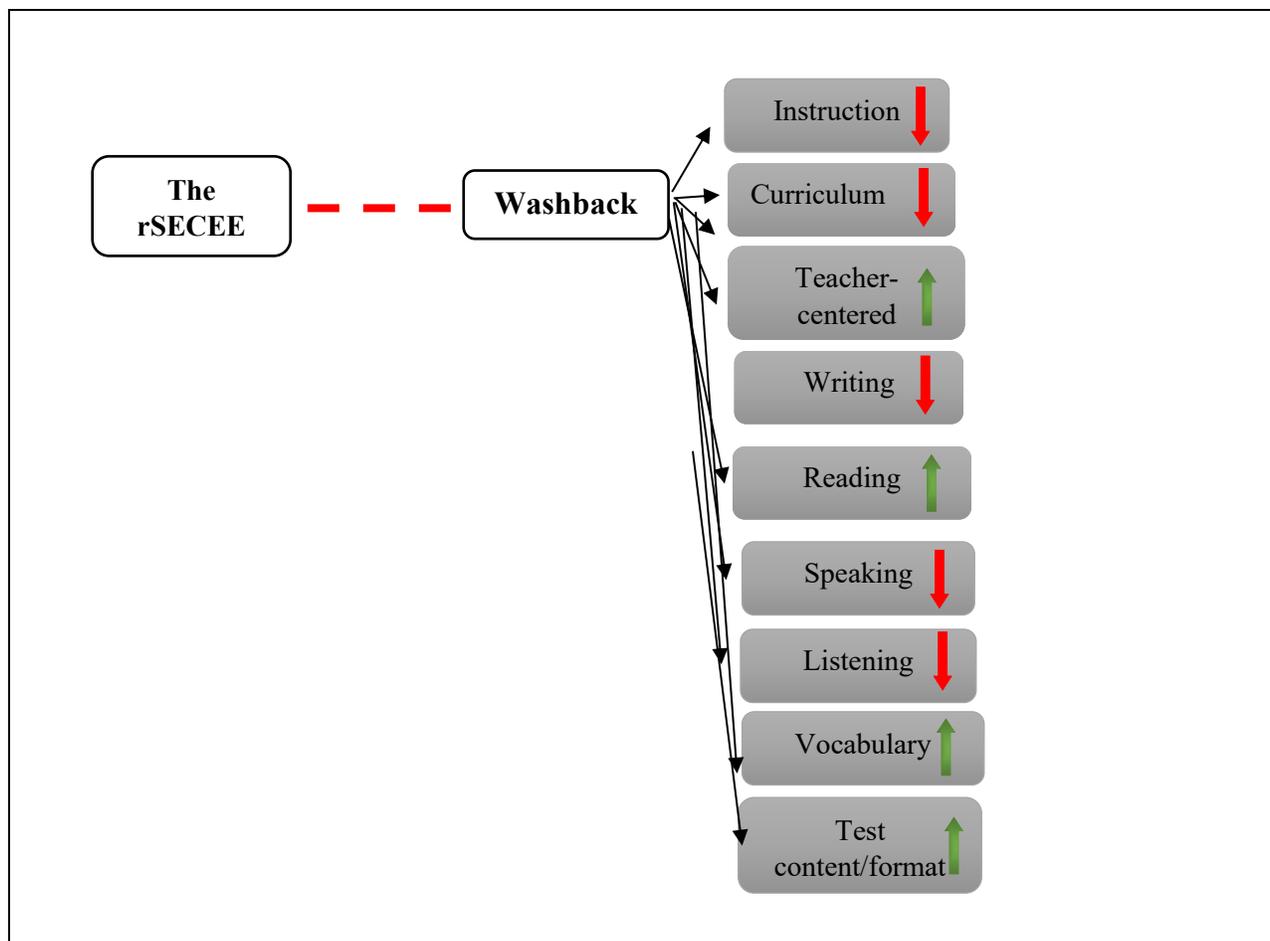


Figure 5.1: Hypothesised Washback of the rSECEE on Grade 12 EFL Classrooms  
Source: adapted from Wang (2010).

These presuppositions are depicted in Figure 5.1. The red dotted line indicates the hypothesised relationship between the rSECEE and its washback on the Grade 12 EFL classrooms. The green arrows show the increases that were expected to occur, while the red arrows indicate where decreases were expected to occur.

### 5.3. Research Philosophy

It is important to explain the research orientation in this study, as the reliability of any study is improved when researchers are explicit about both the process of their research and their positioning in relation to their study (King, 1998). I approached this study from within the pragmatic philosophic tradition.

It is worth noting that a philosophical debate known as the ‘paradigm wars’ surfaced at around the late 60s and early 70s (Johnson & Onwuegbuzie, 2004). This philosophical debate “left educational research divided between two competing methods: the scientific empirical tradition [quantitative research], and the naturalistic phenomenological mode [qualitative research]” (Burns, 1997, p. 3). It may be argued this debate led to a third new approach known as mixed methods, which holds the view that the main determinant of the epistemology,<sup>17</sup> ontology<sup>18</sup> and methodology<sup>19</sup> one adopts is the research question (Saunders, Saunders, Lewis, & Thornhill, 2011). In other words, the pragmatic approach positions “the research problem” as central, and employs all approaches to understand the problem (Creswell, 2003, p. 11). With this in mind, Tashakkori and Teddlie (1998) advise that it is more advantageous for a researcher conducting a study to think of the adopted philosophy as a continuum rather than opposing positions. As such, the pragmatic approach offers the opportunity for “multiple methods, different world views, and different assumptions, as well as different forms of data collection and analysis” in one single study (Creswell, 2003, p. 11).

Mixed methods have become increasingly popular, primarily because it is perceived to provide an epistemological basis for mixing approaches and methods (Creswell & Plano Clark, 2007; Onwuegbuzie, Johnson & Collins, 2009; Teddlie & Tashakkori, 2009). Along these lines, Tashakkori and Teddlie (1998) advise the researcher that he/she ought to study what interests and is of value to him/her and to consider all the various methods that are deemed to be appropriate.

---

<sup>17</sup> The process of thinking, the relationship between what we know and what we see (Guba & Lincoln, 2005).

<sup>18</sup> The assumptions we make about the way in which the world works.

<sup>19</sup> The process of how we seek out new knowledge. The principles of our inquiry and how inquiry should proceed (Schwartz, 2007, p. 190).

#### 5.4. Research Design

As mentioned in Chapter One, the overall purpose of this study was to investigate the relationship between the degree of alignment among the components of an educational system (i.e., standards, curriculum, and testing) and the washback on Grade 12 classroom teaching and learning. A comprehensive review of the literature, and the need for a rigorous and comprehensive approach to better understand how the rSECEE operates at the classroom level, suggests that the research should include the collection and analysis of both quantitative and qualitative data. Therefore, the study employed a mixed-methods explanatory sequential design (Creswell, 2015) research design (Creswell, 2015) consisting of two phases.

Phase I (presented in Chapter Six) was an alignment study that assessed the degree of alignment between the rSECEE and Libya's EFL content standards. Subsequently, Phase II (presented in Chapter Seven) employed qualitative methods to understand the washback of the rSECEE on the Grade 12 Libyan EFL language classroom. Data was collected through two questionnaires, semi-structured interviews and observations. The first questionnaire (see Appendix E) was designed to elicit information about the Libyan EFL teachers' accounts and concerns regarding the rSECEE. The use of individual interviews, and observations provided the researcher with the opportunity to probe deeper into both the participants' accounts and practices. Findings from the degree of alignment between the rSECEE and Libya's EFL content standards constituted the baseline evidence against which Phase II findings were assessed. The multiple measures involved helped to validate the quantitative and qualitative findings, and thus, the findings were "more convincing and accurate" (Yin, 2009, p. 116). In essence, employing different data sources in a single study acts as a method of data triangulation (Patton, 2002). Patton (2002) contends that "triangulation strengthens a study by combining methods. This can

mean using several kinds of methods or data, including using both quantitative and qualitative approaches” (p. 247). Thus, the use of varied data sources and multiple methods was employed in this study to validate and cross-check the findings (Creswell & Miller, 2000; Creswell & Plano Clark, 2007). Multiple methods were also employed to help note any discrepancies in the questionnaire data and were cross-checked against the data gathered from observations and interviews. Overall, engaging multiple methods in this study “lead to more valid, reliable and diverse construction of realities” (Golafshani, 2003, p. 604).

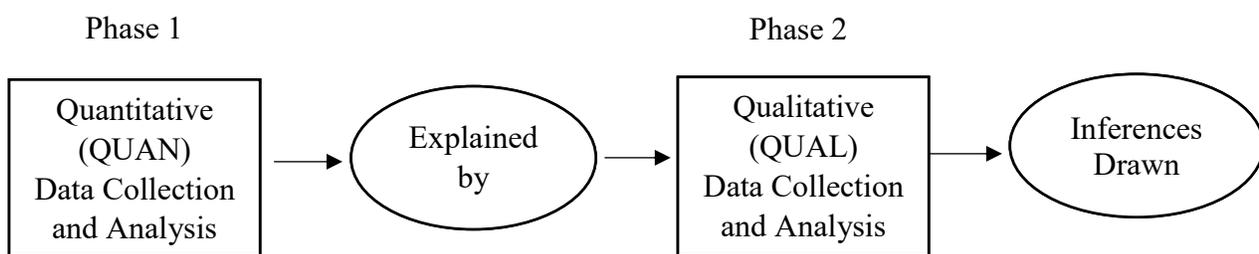


Figure 5.2: Explanatory Sequential Design

Source: Creswell (2015, p. 39)

The research design, with its phases, is depicted in Figure 5.2. The integration of findings is key characteristics of mixed methods research (Creswell, 2015). In the present study, equal weighting was given to both quantitative and qualitative data collection and analysis, and the results of each Phase were integrated. The design was selected as the one that would allow for the most useful type and amount of information to be collected in order to best answer the research questions.

#### 5.4.1. Justification of the Research Design

Creswell and Plano Clark (2007) define mixed-methods (MM) research as an approach that guides the collection and analysis of both qualitative and quantitative data in a study. MM

developed its uniqueness in response to the incommensurability thesis of the paradigm wars (Gorard, 2004). MM researchers do not consider quantitative and qualitative research as two opposing poles. Instead, they regard quantitative and qualitative research as a continuum along which different research approaches fall (Tashakkori & Teddlie, 1998; Teddlie & Tashakkori, 2009, Turner, 2014). In essence, the logic of inquiry for MM research involves the employment of induction (discovery of patterns), deduction (testing of theories and hypotheses), and abduction (uncovering and relying on the best of a set of explanations for understanding one's results) and depends on the best set of explanations for understanding one's results (Johnson & Onwuegbuzie, 2004, p. 17).

Furthermore, the MM research approaches allow the research questions to determine the data collection and analysis (Mackenzie & Knipe, 2006; Turner, 2014). The importance of MM research approaches is that it allows researchers to mix aspects of the qualitative and quantitative paradigms in all or most of the methodological steps in the design (Creswell & Plano Clark, 2007, 2011). The main aim of MM research is to draw from the strengths of each while mitigating the limitations of each in one research project. Because of the increased use of MM research, it has been identified as a "key element in the improvement of social science" (Gorard, 2004, p. 7). Gorard (2004) further emphasizes that MM research "can lead to less waste of potentially useful information, creates researchers with an increased ability to make appropriate criticisms of all types of research" (p. 7). The reasons for using MM design and approaches in this study are summarised in Table 5.1.

Table 5.1

*Reasons for Using Mixed Methods Design and Approaches (Source Adapted from Bryman 2006; and Saunders et al. (2011).*

| <b>Reasons</b>          | <b>Explanation</b>   |
|-------------------------|--|
| Triangulation           | Use of two or more sources of data/data collection methods to validate research results within a study.  |
| Facilitation            | Use of one data collection method or research strategy to aid research using a different form of data collection method or research strategy within a study.                                   |
| Complementarity         | Use of two or more research strategies to complement the weakness of each individual research approach (e.g. quantitative methods complement the employed qualitative methods, and vice versa) |
| Aid interpretation      | Use of both approaches to data collection and analysis to help interpretation and explain the relationship between the sets of data.   |
| Study different aspects | Quantitative approaches can be used to investigate the macro aspects of a phenomenon and qualitative for its micro aspects.  |
| Solving a puzzle        | Use of a different data collection method when the initial method reveals inadequate data.   |

A further justification for using MM research is its growing popularity in research in language testing and assessment in general (see, Cheng & Fox, 2013) and washback research in particular see for example, Andrews et al. (2002), Burrows (2004), Baker (2010), Cheng (2001, 2003, 2005), Green (2007), Erfani (2013), Turner (2005, 2008, 2009), Wall (2005), Wang (2010) and Watanabe (2004). On the value of combining both quantitative and qualitative approaches in washback research, Sturman (1996) emphasizes that the two different types of data elicited yield different types of information. Thus, the value of open-ended comments in questionnaires allow the participants' "depth of feeling to be expressed" (Sturman, 1996, p. 350), while the quantitative data offer "the opportunity to see how representative the written comments are and

whether these comments are distributed randomly through the sample” (Sturman, 1996, p. 350).

Consequently, using the two forms of data can provide a balance between the evidence.

Moreover, owing to the strengths briefly discussed above, and the importance of examining “what teachers think or talk about ... what teachers actually do in the classroom and what actual changes have been made consciously and subconsciously in the classroom” (Cheng, 2001, p. 24), the majority of washback studies have combined both quantitative and qualitative approaches.

Many researchers, including Mullen (2009), Tan (2009), Turner (2006, 2009), Wall and Alderson (1993), and Wang (2010), argue that their research design and analysis have been well-informed by the principles of the MM research approach. Similarly, Wall (1999) further argues that washback research questions are better answered with an MM research design rather than through a single method. This view is supported by Chen (2002) and Turner (2005, 2008) who assert that qualitative and quantitative methods can be profitably used together when investigating washback. MM research can also be advantageous as it is able to “help respond to certain types of questions, especially those having to do with classroom context” (Turner, 2009, p. 108). In this regard and given the growing recognition of the role and benefits of MM approaches in washback studies, this approach was deemed to be the best suited for the study.

## **5.5. Methods**

This research received clearance from Carleton University Research Ethics Board which follows the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (see Appendix F for Ethics Clearance Certificate. I did not commence the data collection until all the participants had signed consent forms stating their willingness to participate in this research.

### 5.5.1. Participants

The study elicited data from alignment review panel members, policy-makers, Libyan Grade 12 EFL teachers, and Grade 12 students. The consideration of different groups of stake-holders in relation to the same phenomenon can allow the researcher to gain a range of perspectives (Darlington, & Scott, 2002).

In Phase I, A panel of educators and assessment and curriculum experts were recruited and trained. In accordance with former research that used the Webb model, such as Roach et al. (2008), panel members were trained to use an analytic and heuristics process to examine the degree of alignment between the Libyan EFL content standards and the rSECEE.

Policy-makers responsible for the implementation of the reform policy regarding the Grade 12 curriculum and assessment were interviewed as a follow up triangulation strategy in Phase I of the current study.

EFL Libyan teachers were considered as participants because it has been established by washback research, including Alderson and Hamp-Lyons (1996) and Spratt (2005) that the teacher can be solely responsible for bringing about washback, whether positive or negative. Evidently, teachers are the most visible stakeholders affected by washback. They are “the ‘front-line’ conduits for the washback process related to instruction” (Bailey, 1999b, p. 17). Furthermore, the importance of teachers and teaching in washback processes was strongly emphasized by Alderson and Wall (1993) in several of their washback hypotheses (1, 3, 4, 7, 9 and 11), and as pointed out by Cheng (2001), to understand washback “we need to ask teachers and watch them” (p. 20). Moreover, it can be argued that in contexts such as Libya where teachers are responsible for high-stakes student assessments, it is fundamental for teachers to be fully aware of the test requirements and assessment criteria (Tan & Turner, 2015). Thus, the

study analysed teachers' understanding of the assessment criteria and their accounts of teaching and the impact of testing.

It is worth noting that eliciting teachers' beliefs may be a difficult endeavour, because personal philosophies may be largely intuitive and subconscious; teachers may not even be able to explain or put them into words. Relatedly is the concern of self-worth and self-esteem; either subconsciously or not, "teachers may wish to promote a particular image of themselves" (Donaghue, 2003, p. 345). Additionally, "there is often a difference between espoused theory (theory claimed by a participant) and theory in action (what a participant does in the classroom)" (Breen et al., 2001, p.345). However, as demonstrated by Woods (1996), teachers' beliefs about learning and classroom pedagogies can be elicited through questioning and searching, or through observing classroom practice and discussing it thereafter.

Qualitative feedback from students on the operation of tests can be seen as fundamental in the Libyan context. As mentioned earlier, all tests have consequences for the test takers and the educational institutions which base their decisions on test scores (Fulcher, 1997). Furthermore, if a test is not viewed as fair by the test-takers, the role of the test may be compromised (Fulcher, 1997). Evidently, learners are the participants whose lives are most directly influenced by any form of testing, including language testing. Although learners' futures are affected by the consequences of tests, they have not been as extensively studied as teachers (Green, 2013; Wall, 1999). Therefore, it is worthwhile distinguishing learners from other stakeholders since the washback process can have a direct impact on language learning (or non-learning). However, the influences on other stakeholders will also influence the efforts to promote language learning (Baily, 1999b).

In Phase 2, a questionnaire and focus group interviews were employed to elicit feedback from test takers. Focus group interviews with learners were mainly used as a triangulation strategy for Phase II.

## **5.6. Sampling**

This study employed a purposive sampling technique because any research sample size depends on the objectives of the research, the research questions, and, more specifically, what the researcher needs to uncover within the available sources (Patton, 2002).

### **5.6.1. Phase I**

A purposive sample was employed to select the alignment review panel members. the criteria for selection and sample size are described in Chapter Six.

### **5.6.2. Phase II**

Phase II qualitative methods using two questionnaires, observations, semi-structured interviews, and focus group interviews, which lent itself to a purposive/judgemental sampling technique. According to Tashakkori and Teddlie (2003) purposive sampling involves the selection of cases “based on a specific purpose rather than randomly” (p. 713).

Purposive/judgemental sampling was employed because cases need to be selected that enabled the research questions to be answered and that drew out rich information. The criteria of selection are discussed further in Chapter Six.

## **5.7. Data Collection**

As illustrated in Figure 5.2, the study employed sequential timing in which the methods were implemented in two distinct phases. Thus, the collection and analysis of each data set was undertaken before the collection of the next data set. The study started with the collection and

analysis of the quantitative data and then moved onto the collection and analysis of the qualitative data.

### **5.7.1. Phase I**

#### **5.7.1.1. Document Analysis**

Taking into account Cheng's (1998) emphasis on the importance of investigating the level of compatibility between the curricular objectives and the related high-stakes test, a document analysis took place prior to the alignment process. Documentary data for Phase I included the Libyan EFL content standards that the researcher accessed from ministry officials, and an analysis of the assigned Grade 12 EFL content objectives, Grade 12 EFL student textbooks, workbooks and the teacher's book<sup>20</sup>. These documents were analysed in order to reveal: the main aim of EFL teaching in Libya; the main aim of the revised version of SECEE curriculum; the intended washback it expects to exert on Libyan EFL classrooms; and the means by which policy-makers intend to use it in order to ensure positive washback. In accordance with Hamp Lyons (1998), the textbook analysis investigated whether or not the textbooks consisted of test-taking strategies, language structures, and tasks that promoted test content and type. The analysis provided the researcher with information that enhanced the overall understanding of the Libyan reform policy and to establish if factors other than curricula objectives might have been in operation and impacted both teaching and learning in the Grade 12 EFL Libyan classroom. The results regarding the underlying principles underpinning the rSECEE and its intended washback were validated in the follow-up interviews with policy-makers at the Ministry of Education.

---

<sup>20</sup> The Teacher's Book provides the teacher with a comprehensive guide to using the course.

### **5.7.1.2. Alignment Analysis Process**

An approach recommended by Webb (1997, 1999) was employed to examine the alignment of the rSECEE with Libya's EFL curricular content standards. This procedure combined "qualitative expert judgments and quantified coding for evaluating the alignment of standards and assessments" (Flowers et al., 2006, p.202). The product of the analysis was statistics that indicate the degree of alignment between Libya's EFL curricular content standards and the content in the rSECEE.

The overall alignment analysis process raised the following questions:

1. Is the content of the test consistent with the test construct?;
2. Does the test measure the intended objectives?; and
3. Does the test content align with the intended curriculum?

An important goal of Phase I was to develop a valid procedure for performing alignment analysis of the Libya's EFL standards and the rSECEE. To the best of my knowledge, this is the first time that such a comprehensive analysis has occurred using the alignment criteria discussed in Chapter Three. A more detailed description of the data collection procedure that took place for the Phase I study is provided in Chapter Six along with its findings.

### **5.7.2. Phase II**

The main problem in relation to investigating a test that has been in operation for a period of time is the lack of baseline data to capture the situation prior to the introduction of the test. In other words, "[t]he investigator arrives 'after the fact'...and tries to determine causal relationships" (Merriam, 1988, p. 7). In this case, the curriculum and testing process have been in operation for a number of academic years. Consequently, it was not possible to implement a baseline study in which the characteristics of the target context could be identified. In order to

manage this issue, the teachers' accounts on how the former curriculum and testing operated at the classroom level were elicited. Such accounts were obtained through the semi-structured interviews in order to get an idea about the teaching prior to the revised test administration. This was then compared to teaching after the introduction of the revised system.

#### **5.7.2.1. Instruments**

A brief description of each tool is given below. However, a more detailed description of each tool and the implemented procedures are presented in Chapter Seven (see also appendix R) along with the findings and a discussion of the findings.

- **Questionnaires**

Questionnaires and interviews are employed in washback studies in order to gain not only information about the participants' demographics, but also in-depth insights into the participants' attitudes, reported practices, and perspectives on language teaching and learning (Cheng, 1998, 2005; Erfani, 2013; Qi, 2005; Tan, 2009; Turner, 2005, 2009; Wall, 2005). Washback researchers such as Cheng (2004), Wall and Alderson (1993), and Watanabe (1996), view questionnaires as valuable tools for providing a general picture of how teachers and learners behave in the contexts of their study. In addition, interviews provide researchers with the opportunity to detect and explain "why teachers do what they do, what they understand about underlying principles... and what they believe to be effective means of teaching and learning (Wall & Alderson, 1993, p. 62).

In Phase II of this study, three Libyan Grade 12 EFL teachers completed the teacher questionnaire, and 15 Grade 12 Libyan students completed the student questionnaire. The questionnaires (see Appendix G for student questionnaire) provided the researcher with an overview of two key stake-holders and their accounts of language teaching and learning and

testing. In particular, the teacher questionnaire provided background information about the teachers, their perceptions about the rSECEE, their experience of teaching Grade 12 English curriculum, and how they perceived its impact on classroom teaching and learning. The insiders' views were fundamental to understanding teachers' changes (if any) after the introduction of the revised Grade 12 EFL curriculum and SECEE. Data was in the form of items and extended written responses. Similarly, the student questionnaire provided background information about the students, their experience of learning English, the perceptions of in-class test preparation practices, and any challenges experienced whilst learning and preparing for the test. These instruments helped the researcher to understand a system from the perspectives of those involved with it (Kaplan & Maxwell, 2005).

The construction and administration of the teacher questionnaire took place after the results of Phase I became available, as these informed the construction of the questionnaires. Teacher questionnaire items were similar to the questionnaires employed in Abdulhamid's (2011, 2013) studies, which had already been tested and validated. To further enhance the questionnaire's construct validity and reliability it was piloted with a sample of three Libyan EFL teachers who did not participate in this study. The piloting procedure helped to check the clarity of the instructions, the questionnaire items, and the time allocated for the questionnaire to be completed (Cohen, Manion & Morrison, 2000). Importantly, the researcher's experience as a high school teacher, a full acquaintanceship with the context of study, findings from previous research conducted in Libya, interviews with education officials, and the researcher's personal contacts with Grade 12 teachers and students, all played a role in enhancing the content validity of the questionnaire. As noted by Cheng (1998), the researcher's insider knowledge is important in this context.

The student questionnaire was not constructed until the researcher completed the collection of data from the classroom observations and the individual teacher interviews. It was believed that the results from the observations and interviews would enhance the construct validity of the student questionnaire. In addition, the student questionnaire was piloted for the same reasons as the teacher questionnaire. The questionnaire was piloted with a sample of five Grade 12 students.

### **Observations**

Teachers who participated in the study were observed in their classrooms because of its importance in understanding the nature of washback (Alderson & Wall, 1993; Alderson & Hamp-Lyons, 1996; Bailey, 1999; Cheng, 2001, 2008; Wall & Anderson, 1993; Watanabe, 2004). The use of observation as a tool for eliciting data in washback studies is recommended by Turner (2001), Wall (1999), Wall and Alderson (1993), and Watanabe (2004) as a method for contextualising, confirming, or validating data from questionnaires and interviews. Furthermore, Wall and Alderson (1993) argue that without observation it is unlikely that inconsistencies between what teachers perceive they are doing and what they are actually doing would be revealed. Similarly, Bailey (1999b) contends that if the core of washback is related to the effects of tests on teaching and learning, it would seem necessary to document those effects by both asking about and watching teaching and learning. Thus, it can be argued that observation is a necessary component in understanding washback. Overall, the aims of the classroom observations were to help determine what and how teachers teach, to examine the relationship between the rSECEE and the EFL teaching activities, and to “probe deeply and to analyze intensively the multifarious phenomena that constitute the life cycle of the unit with a view to establishing generalizations about the wider population to which that unit belongs to” (Cohen & Manion, 1989, p. 124).

A series of presuppositions were proposed for the observation phase, based on the teachers' responses to the questionnaire, and the assumption that whatever the teachers reported would be observable in their classroom behaviour. Three EFL teachers in five classes each (in total 15 classes i.e., 675 minutes) were observed. Previously, the researcher observed two classes in order to pilot the method and tools of observation. After each observation, the researcher asked the teachers about their objectives behind certain activities and episodes that took place within the lesson. These post-observation chats took place in Arabic and were later translated into English. A further description of the observation process and the challenges that the researcher faced are presented in Chapter Seven, along with the Phase II findings. It is worth noting that prior to the observations the three teachers were informed of the main purpose of the research but were not asked about their attitudes towards the rSECEE, in case their awareness was raised, and hence "pollute" their teaching (Watanabe, 2004, p. 133). The teacher observational data lead to a number of conclusions that guided the interviews through the development of further hypotheses.

- **Semi-Structured Interviews**

As mentioned, Phase II investigated teachers' and students' perspectives of the rSECEE, and any possible reported impact it may have on them. A semi-structured interview was deemed to be a suitable instrument for yielding data and meeting the needs of the study. As argued by Dornyei (2007), a semi-structured interview is "suitable for cases when the researcher has a good enough overview of the phenomenon or domain in question and is able to develop broad questions about the topic in advance but does not want to use ready-made responses categories that would limit the depth and breadth of the respondent's story" (p. 136). More specifically, the purpose of conducting the interviews was two-fold. The first reason was to elicit further information on the questionnaire responses and teacher perceptions and themes. The second was

to understand more about teachers' behaviours and classroom practices observed during the observations. The interviews were conducted with the teachers who were observed and were conducted after the results from the teacher questionnaire and observations became clear. During the interviews, the teachers were asked questions that the researcher had prepared in advance. However, the researcher also asked other questions to probe points that came up during the interview. The interview questions (see Appendix H) were designed with the intention of exploring teachers' accounts and beliefs about the influence of the rSECEE on their teaching, and if they had made any changes in their approach to teaching as a result of the examination. Both the observation and interview questions were based on a number of Alderson and Wall's (1993, p. 120-12) hypotheses:

1. A test will influence what teachers teach;
2. A test will influence how teachers teach; and
3. Tests will have washback effects for some learners and some teachers, but not for others.

A more detailed description of the interview process is presented in Chapter Seven.

- **Focus Group Interviews with Students**

A focus group is defined as a “group of individuals selected and assembled by researchers to discuss and comment on, from personal experience, the topic that is subject of research” (Powell, Single & Lloyd, 1996, p. 499). It is important to note that focus groups differ from interviews in that they rely on communication within the group based on issues that are provided by the researcher (Morgan, 1997). Furthermore, rather than reaching a consensus, focus groups have the advantage of bringing to the floor different views on a particular matter (Kvale, 2007; Morgan, 1988). Arguably, learners' attitudes, “feelings, and beliefs may be partially independent of a group or its social setting” but are more likely to surface through “the social gathering and the

interaction which being in a focus group entails” (Gibbs, 1997, p. 7). Focus group interviews were chosen as an avenue for data collection as they are an easy way to collect data and take less time in comparison to other qualitative methods such as observation and interviews (Al-Hamdan & Anthony, 2010). In addition, “[i]n depth, conversational exchanges between participants and the moderator offer[ed] an opportunity” for the researcher “to hear not only what participants are thinking and feeling but also the details about circumstances through which meaning has been constructed” (Stalmeijer, McNaughton, & Van Mook, 2014, p. 926). This study used the focus group interviews to elicit feedback from test-takers about their: accounts of the rSECEE; experience of learning English; reported impact of the rSECEE on their learning; reported challenges whilst learning and preparing for the test; and out-of-class English learning practices. In addition, the study elicited feedback regarding test transparency and issues related to test administration (for example, whether students felt exhausted during the test, or if they were able to deliver their best performance) because test-takers ought to be familiar with test tasks before sitting the actual test. It was noted that the level of test-takers’ familiarity with test demands may affect the manner in which the test task is approached, and, consequently, affect test performance (Weir, 2005).

The focus group themes were based on students’ responses from the questionnaire, and addressed three of Alderson and Wall’s (1993, p. 120-122) hypotheses:

1. A test will influence what learners learn;
2. A test will influence how learners learn; and
3. Tests will have washback effects for some learners, but not for others.

A thorough description of the observation process is presented in Chapter Seven. How each method aims to answer the research questions is summarised in Figure 5.4.

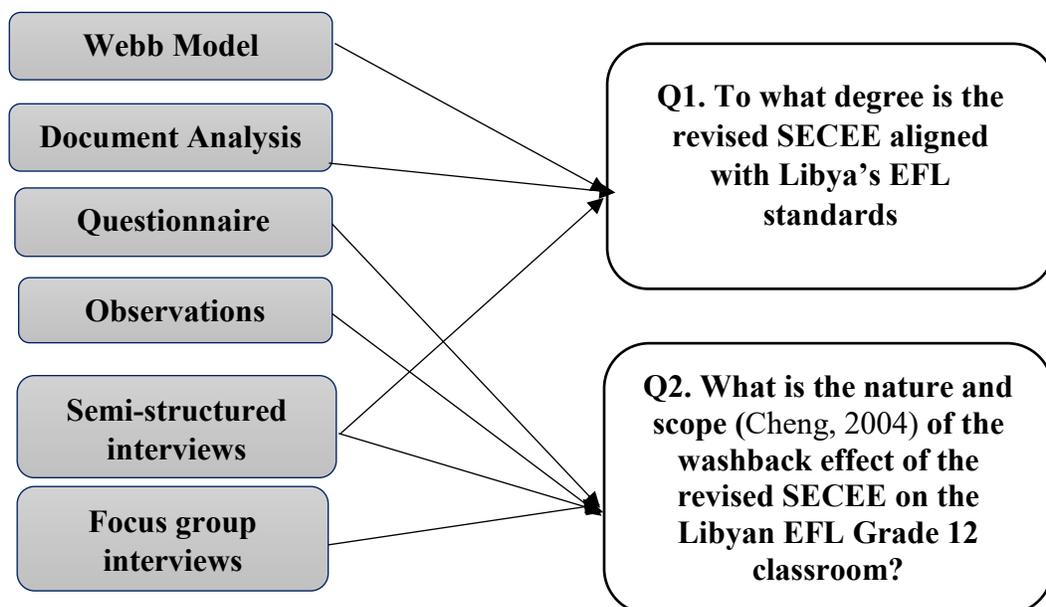


Figure 5.3: Methods and Research Questions

## 5.8. Analysis

A sequential data analysis approach was employed in this study. The quantitative data in Phase I was analysed first, and the results were used to inform the subsequent qualitative Phase (Phase II). A more detailed description of the analysis procedures for Phases I and II are provided in Chapters Six and Seven respectively. Finally, both the quantitative and qualitative strands of analysis were synthesized and merged into one overall interpretation (Creswell, 2000).

## 5.9. Reliability and Trustworthiness

The quantitative component of the study was validated by piloting the tools and procedures employed in the alignment analysis process, while the teacher and student questionnaire were piloted in order to improve their reliability. In contrast, the qualitative component's data trustworthiness was established by member checking, providing rich and thick descriptions, and triangulation (Brown, 2001; Creswell & Miller, 2000, Creswell & Plano Clark, 2007). In brief, member checking or "respondent validation" includes "techniques in which the investigator's

account is compared with those of the research subjects to establish the level of correspondence between the two sets” (Mays & Pope, 2000, p. 51). In addition, as recommended by Yin (2009) key informants reviewed that data and their interpretations. Any corrections made by key informants enhanced the accuracy of the information as well as “identif[ied] a range of competing perspectives” (Baskarada, 2014, p. 7). Thick descriptions involved presenting the data in adequate richness and detail to persuade the reader of the validity and sufficiency of the findings, and to achieve transferability (Stalmeijer, McNaughton & Van Mook, 2014). Transferability was also achieved in this study by relating the appropriate literature and previous research findings to the interpretations of the results, as this helps readers to compare the degree of transferability and suitability for their context.

The “key aspect of qualitative research is that its objective is not to produce findings that are capable of general application, but rather to produce results that resonate” (Senior, 2006, p. 16). Senior’s use of the word resonate refers to the ability of the research findings to be meaningful and relevant to those who experience them. Furthermore, readers are more likely to be convinced of the reliability and validity of the research conclusions based on the “clarity and comprehensiveness of evidence” provided (Bazeley, 2012, p. 151).

Finally, triangulation obtained “different but complementary data on the same topic” (Morse, 1991, p. 122) in order to better understand the topic. Similar to washback studies (such as Cheng 1997, 1998; Turner 2008, 2009; Wall, 1999; Wall and Anderson, 1993) that employed triangulation techniques and Bailey’s (1996) emphasis on incorporating methodological triangulation in washback research, both data (use of varied data sources) and methodological (use of multiple methods to study a problem) triangulation were employed. Utilising a number of data sources allowed for “the development of converging lines of inquiry (Baskarada, 2014, p.

12), and the inclusion of multiple cases played a role in enhancing the “external validity” and “generalizability” of the findings (Merriam & Tisdell, 2016, p. 40).

In addition, thick and rigorous descriptions of the study’s context was provided as this approach helped to enhance the validity and interpretations of the results. The results are used to promote insights and enhance our knowledge and understanding of high-stakes testing and its possible effects on teaching and learning, rather than providing causality and proof (Bailey, 1999). Furthermore, the findings are believed to have face<sup>21</sup> validity. They were believable to other Grade 12 EFL teachers who work or had worked at the schools where the study was conducted, to EFL language inspectors, and even to teachers who have children undertaking either the rBECCE or rSECEE. The findings are credible because they are reliable and the description of the Grade 12 EFL teachers and their language classroom are recognisable and resonate with Libyan EFL teachers. The following chapter presents a more detailed description of the data collection and analysis procedures for Phase I, along with the results, and discussion of the results.

---

<sup>21</sup> Face validity is defined as the appropriateness, sensibility or relevance of the results. It is also degree to which participants within in a system view the findings of this research as relevant to the context in which the study took place.

## CHAPTER VII

### Phase I

As discussed earlier in this dissertation, the evaluation of the degree of alignment amongst assessment, curricula and standards is receiving increasing attention in the educational field (Li & Sireci, 2004). Given the context of this study and the increasing assessment demands that are associated with Libya's recent reform policy, it is critical to provide evidence about the degree to which the rSECEE is compatible with Libyan EFL content standards and educational goals.

As noted in Chapter Three, alignment of large-scale assessment and content standards is among the evidence gathered in the validation of test score interpretations (Forte, 2016). In addition, the current edition of *The Standards for Testing* (AREA, APA, NCME, 2014) emphasizes the importance to test users of gathering validity evidence to support test score interpretation and use. As quoted in *The Standards for Testing* (AREA, APA, NCME, 2014), "an analysis of the relationship between the content of a test and the construct it is intended to measure" (p.15) is essential. Furthermore, a "Clear articulation of each intended test score interpretation for a specified use should be set forth, and appropriate validity evidence in support of each intended interpretation should be provided" (p.23). It can be further contended that as in former editions, the 2014 *Standards* highlight for test evaluators five areas for evidence-gathering in a validity evaluation process. These five sources of evidence, as explained by Forte (2016), are "neither types of validity nor discrete boxes on a checklist" (p.6). Instead, they are areas where test evaluators and users attempt to answer questions with regards to validity. Forte (2016) goes on to stress that alignment research is of critical importance, because "without alignment evidence—from multiple sources—it would be impossible to interpret assessment scores in relation to the standards on which those assessments are meant to be based" (p.6).

In this context and given the growing recognition of the role and benefits of alignment research, we can see the importance of the Phase I study, which addressed the first research question of this dissertation, namely:

**To what degree is the rSECEE aligned with Libya’s EFL content standards?**

It is important to note from the outset of this chapter that different countries use different labels or terms for identifying levels of educational expectations. In this dissertation, the conventions of content standards and “content objectives” are employed to describe two levels of expectations that students are expected to know and achieve. It is assumed that all the Libyan EFL content objectives under the standards span the content knowledge expressed in the Libyan EFL content standards. In other words, the Grade 12 Libyan EFL content objectives are treated as more detailed descriptors of what students are expected to know and do in order to demonstrate their attainment of the standards; hence, they are a more detailed specification of the standards. In addition, the term *test item* is used to represent the behaviour that is elicited from a student/ test taker.

In this chapter, a more detailed description of the data collection and analysis procedures for Phases I is provided.

## **6.1. Method**

### **6.1.1 Participants**

A purposive sample was employed to select the alignment review panel members. To the best of my knowledge there is no recommended number of experts for conducting alignment analyses (FitzPatrick, Hawboldt, Doyle, & Genge, 2015), so a sample of four ( $n = 4$ ) was deemed reasonable (Martone & Sireci, 2009). To obtain a representative sample, alignment review participants were recruited according to certain parameters. They were language teachers

who have an extensive understanding of measurement and testing and who have experience with instruction and assessment involving English as a second and/or foreign language. The majority of the panel members had experience and were familiar with rating language proficiency tests, such as IELTS, CAEL, TOEIC, and TOEFL; all of which are national or international high-stakes tests. Two of the panel members were assessment specialists with a research focus on test reliability, validity, high-stakes testing impact, and diagnostic assessment. One of the two assessment specialists is the primary developer of several high-stakes language tests, including the CAEL, and has numerous publications in the field of language assessment. Importantly, panel members were not involved in the development of either rSECEE, or Libya's EFL content standards. The alignment panel members were contacted and invited to participate via email.

### **6.1.2. Policy-Makers**

As mentioned in Chapter Five, policy-makers responsible for the Grade 12 curricular implementation were interviewed. Specifically, ministry officials were contacted to verify the Libyan EFL content standards and the Grade 12 content objectives. In addition, a Libyan test developer who had participated in developing the rSECEE gave his account of the test development and validation process.

## **6.2. Data Collection**

### **6.2.2. Document Analysis**

A document analysis for Phase I was conducted as a first step. As mentioned in Chapter Five, textual data for Phase I included the Libyan EFL content standards that I accessed from ministry officials and an analysis of the assigned Grade 12 EFL content objectives, Grade 12 EFL student textbooks, workbooks and the teacher's book. As recommended by La Marac (2001), the study applied basic discourse analysis of the Libyan Grade 12 EFL content objectives. La Marca

(2001) states that the curriculum has to be a part of the alignment analysis because content standards only provide descriptive statements of any curriculum. The analysis enhanced the researcher's understanding of the Libyan reform policy.

### **6.2.3. Alignment Analysis Process**

The overall steps in the alignment analysis process included:

1. Identifying criteria and acceptable levels;
2. Identifying the expectations of the Grade 12 EFL Libyan subject;
3. Developing the coding matrix for the content domain;
4. Training alignment review panel members;
5. Having the alignment review panel code test items in relation to objectives;
6. Entering data codes onto a spreadsheet;
7. Analysing data;
8. Preparing summary data tables; and
9. Reporting results.

In the accordance with other alignment studies (such as Herman et al. 2007; Roach et al. 2005, 2008; Webb et al. 2002) the main role of the alignment review panel members in studies employing Webb's alignment model, including this study, was to:

1. Arrive at an agreement on the DOK level rating for each objective in a state's/nation's content standards;
2. Classify the DOK level of each test task or item; and
3. Associate the one/two objectives from the content standards to the corresponding test items.

The study used the rSECEE for the academic year 2014/2015. A panel of four subject matter experts (SMEs) was trained. It is worth noting that the alignment process was piloted twice with two SMEs to improve the “data collection plans with respect to both the content of the data and the procedures to be followed” (Yin, 2009, p. 92).

#### **6.2.4. Instruments**

##### **6.2.4.1. Libya’s EFL Academic Content Standards for Teaching English in Libyan Secondary Classrooms**

The researcher accessed from a Libyan ministry official the following EFL academic content standards for teaching English in Libyan secondary level classrooms. The purpose of EFL instruction at the secondary level is:

1. To assist the students to manipulate the English language [the students’ linguistic knowledge of the language] as a linguistic system and to have some conscious knowledge of how it works at the level of phonology, morphology, syntax and discourse.
2. To provide the lexical system [students’ knowledge of vocabulary] with words so that the students can discuss topics related to their specialization.
3. To lay the foundations of self-study in English to enable the students to continue learning after school.
4. To help students to achieve satisfaction through using English for personal activities.
5. To encourage the students to appreciate the value of learning English as the most widely used language in the world.
6. To raise awareness of the important role English can play in the general national language and culture in knowledge, technology and experience through translation, and international affairs.

The above Libya's EFL content standards suggest that the expectations set for learning English in Libyan secondary classrooms demand a communicative approach of teaching and learning (Larsen-Freeman, 2000, see section 2.4). In addition, the expectations require different levels of thinking skills on behalf of the students (Dewey 1933; Hmelo & Ferrari, 1997, see also Glossary of terms & section 2.4). The expected levels of thinking appear to range from lower to higher order thinking skills.

The EFL content standards were most clearly articulated in the teacher's book. Every lesson within a unit had its own content objectives. To help organise the content objectives in a more systematic fashion, the researcher listed each objective under its designated language skill (including vocabulary and grammar) and gave it a code (See Appendix I). For example, the content objective "*To develop skills in predicting the content of a text, including vocabulary*" was categorised as a reading skill objective and coded as R1. This strategy facilitated the alignment process because each content objective was given a code that could be easily entered by panel review members rather than them going to the trouble of writing out the full description of the target content objective. The coding of the content objective was also advantageous for the analysis stage. The researcher entered the codes into the Excel spreadsheets, which, in turn, helped the calculation of Webb's (1999) four alignment criteria.

#### **6.2.4.2. The 2014/2015 rSECEE**

The alignment evaluation phase was based on the set rSECEE items for the academic year 2014/2015 which contribute to the scores that are intended to reflect the knowledge and skills defined in the six EFL content standards listed above. As with other versions of the rSECEE, the academic year 2014/2015 version included 60 test items: 25 true and false; 25 multiple choice

(MCQs); and 10 matching items. Three hours were allocated for the examination, and students were not allowed to use dictionaries or translation devices.

#### **6.2.4.3. Semi-Structured Interviews**

The purpose of conducting the interviews with Libyan policy-makers was two-fold. The first reason was to verify the Libyan EFL content standards and objectives of the Libyan testing reform policy. The second was to use the interview data as a follow-up triangulation strategy for Phase I of the present study. The interviews were conducted with two policy-makers. The interviews were conducted in Arabic and then translated into English. During the interviews, the policy-makers were asked questions that the researcher had prepared in advance. However, the researcher also asked other questions to probe points that came up during the interview. The interview questions (see Appendices J & K) were designed with the intention of exploring and validating the results of the degree of alignment between the rSECEE and Libya's EFL content standards. After the two interviews, I used e-mails and phone calls to clarify, extend and validate my understanding. After their clarifications and responses to my questions, I provided on-the-spot validity checks. This was achieved by repeating to each participant what he and she had said during the flow of the interview and to also ensure that I have understood him and her correctly.

#### **6.2.5. Procedure**

##### **6.2.5.1. Identifying Criteria and Acceptable Levels**

Prior to the training session, I used the World Language Cognitive Rigor Matrix (CRM, 2015) for the EFL descriptions of the depth of knowledge (DOK) levels for the Libyan context (see, Appendix L). This Matrix was based on the Hess Cognitive Rigor Matrix (2005, 2009) and Hess' (2013) DOK supports for English Language Learners. The CRM is a model that

superimposed Bloom's Taxonomy<sup>22</sup> with Webb's DOK level (Hess, 2014). Hess, Carlock, Jones, and Walkup (2009) argued that although Bloom's thinking levels and Webb's DOK are similar in terms of the complexity of thought involved, they differ in terms of "scope, application, and intent" (Hess, 2014, p. 1). DOK descriptors in the CRMs offer examples for illustrating "how students might move towards deeper understanding with more complex or abstract content" (Hess, 2014, p.1).

#### **6.2.5.2. Developing the Coding Matrix for the Content Domain**

I developed two coding matrices: one describing the Libyan standards and objectives, and the other listing the test items, for collecting reviewers' data. The first coding matrix enabled the reviewers to evaluate the DOK levels for each SECE content objective (see Appendix M). The second coding matrix listed each test item with columns for reviewers to code the DOK level and content objective (see Appendix N). Besides the two coding matrixes, I developed a package for the reviewers. This package included samples from the Grade 12 EFL textbooks. The samples illustrated to the reviewers the type of text and tasks that Grade 12 students would encounter in these textbooks. Since the student course textbook consisted of eight units, a sample of three units was chosen to represent the Grade 12 EFL textbooks. These three units included tasks that covered the four language skills, as well as vocabulary and grammar. The purpose of preparing the Grade 12 textbook package was two-fold. First, it aimed at providing some context for the reviewers and helped acquaint them with the Libyan EFL textbook materials. Second, it provided

---

<sup>22</sup> Benjamin Bloom (1956) developed a taxonomy to describe the cognitive demand of different learning and assessment tasks. His scheme classified levels of intellectual behaviour that are fundamental in learning (Hess et al. 2009). Bloom's created his taxonomy to categorize the "levels of abstraction of questions that commonly occur in educational setting (Hess et al, 2009, p. 1).

the reviewers with the opportunity to understand to what degree Libyan EFL academic standards have been operationalised within the Grade 12 EFL student textbooks.

### **6.2.5.3. Training Alignment Review Panel Members**

In previous studies that employed Webb's alignment model (such as Herman et al. 2007; Roach et al. 2005, 2008; Webb, 1999), members of the alignment review panel were trained to use an analytic process and heuristics to evaluate the degree of alignment between tests and academic standards. Training, as posited by Webb (1999, 2002), is fundamental to ensure that all review panel members fully understand the standards, nature of alignment tasks and the employed rating scales. It should be emphasized that without an informative training session "the alignment process can easily deteriorate into a non-productive activity" (Bhola et al., 2003, p.27). Against this background, a panel of assessment and curriculum experts ( $n = 4$ ) were trained. The alignment review panel received a short two-hour training session in the use of the analytic process and heuristics to rate the degree of alignment between the rSECEE and Libya's EFL standards. As advised by Webb (1999), an excessive amount of time was not spent on face-to-face training, especially when there were only four SMEs. The reviewers were informed of the overall goals of the study and were sent a detailed description of the training and alignment judging and analysis session via email prior to the actual training day.

In accordance with Bhola et al. (2003) and Webb (1999, 2002), the training session provided the alignment review panel members with a definition of the criteria employed in the study. It also familiarised them with the scales and instrument used to assess the degree of match between test items and standards. At the training session, the panel review members were given additional orientation relevant to the study and its goals. The training also included a review of the four DOK levels of knowledge that are applicable to rSECEE and provided the opportunity for

discussion of the rating criteria. Thus, it guided the panel members' understanding of the DOK rating process (Webb, 2002). As advised by Sireci (1998a), the alignment review panel members were provided with a monetary incentive in order to keep them motivated and on task (Sireci, 1998a).

#### **6.2.5.4. The Alignment Review Panel Coding**

After the training session, the reviewers were asked to assign a DOK level for each rSECEE content objective and enter their codes in coding matrix I. Prior to entering their individual codes and independently completing their ratings, the reviewers were asked to discuss a number of their rated items ( $n=10$ ) and to reach an agreement about the DOK level of the content objectives. The panel members reached a consensus on the DOK for the Libyan EFL content objectives in each of the content domains —reading, writing, speaking and listening. Vocabulary and grammar were also categorised as fixed content domains for the Grade 12 SECEE curriculum. In accordance with Webb (2002), panel members worked together as a group to reach a consensus on the DOK levels for each listed objective, which proved to be advantageous. It provided an opportunity for panelists to discuss the rating criteria, which resulted in a “calibration of panel members’ understanding of DOK rating process” (Roach et al., 2008, p.161). Subsequently, during the alignment session, panelists were also asked to code the first ten test items ( $n = 10$ ) as a group and discuss their results. This step was repeated until the alignment review panel members had reached consistency in their coding and had become familiar with the procedure. Following the calibration process, the panel members assigned a DOK rating to each rSECEE test item, and then completed the coding process by identifying the objective(s) from the standards that corresponded to the test item. Similar to previous alignment studies that

employed Webb's model of alignment, the panel members coded the test items independently, with no interaction between them and entered their codes in the coding matrix II.

After the individual coding process was completed, the panel members selected a sample of items in order to discuss and compare which test items were assigned to which objectives, and which DOK was assigned to the selected test items. Webb (1999) argues that such discussions enhance the rating reliability among the panel members in coding future test items. The panel members were given the option to change their decisions after their discussion with other members. Checks of reviewers' agreement were incorporated at different points until there was sufficient evidence that the reviewers were interpreting the DOK levels, the standards, and the test items in a consistent and systematic way. Figure 6.1 provides an example of coding matrix II and an illustration of how a panel member coded rSECEE test items.

It should be noted that several times during the training session and analysis, reviewers raised questions about the Libyan educational context and how the standards and high-stakes tests were used as well as how they were developed. For instance, deciding on a DOK level for some test items depended on understanding if the students had been exposed to the reading content during the academic year. It would be difficult for an external panel member to synthesize all the essential information about the standards and the rSECEE within the short training and analysis session. Therefore, having the researcher participate in the analysis proved to be helpful and made it possible to answer reviewers' questions on the spot and identify to reviewers the principal piece of knowledge measured by a test item and its purpose. If the reviewers had difficulty deciding between two levels for a content objective (e.g., between a rating of Level 1 or 2), they were advised to choose the higher of the two levels. Furthermore, during the coding

process the reviewers experienced difficulties in remembering and locating objectives that matched the rSECEE test items.

**Please assign each listed test item with the best corresponding DOK level and objective(s) for the sixty test items. If you have difficulty deciding between two objectives for an item, you are advised to choose the two most suitable objectives (please refer to the objective listing handout).**

The SECEE Alignment Review Panel Session  
Reviewers name: [REDACTED]

### Coding Matrix II

| Test item # | DOK Level | Objective(s) |
|-------------|-----------|--------------|
| 1.          | Level 1   | R2           |
| 2.          | Level 1   | V16          |
| 3.          | Level 1   | V6           |
| 4.          | Level 1   | R17/19       |
| 5.          | Level 1   | R10          |
| 6.          | Level 1   | G22          |
| 7.          | Level 1   | G2           |

Figure 6.1: A Sample of a Reviewer's Coding Responses to rSECEE Test Items

After coding for some time, it was difficult for the reviewers to retain the huge number of content objectives. Reviewers reported that they had carefully considered the knowledge required for a student to successfully answer an item. The data were collected in pen and paper form from the reviewers and then entered in an Excel spreadsheet by the researcher. Figure 6.2 illustrates how the criteria measured the relationships between Libya's EFL content standards and rSECEE.

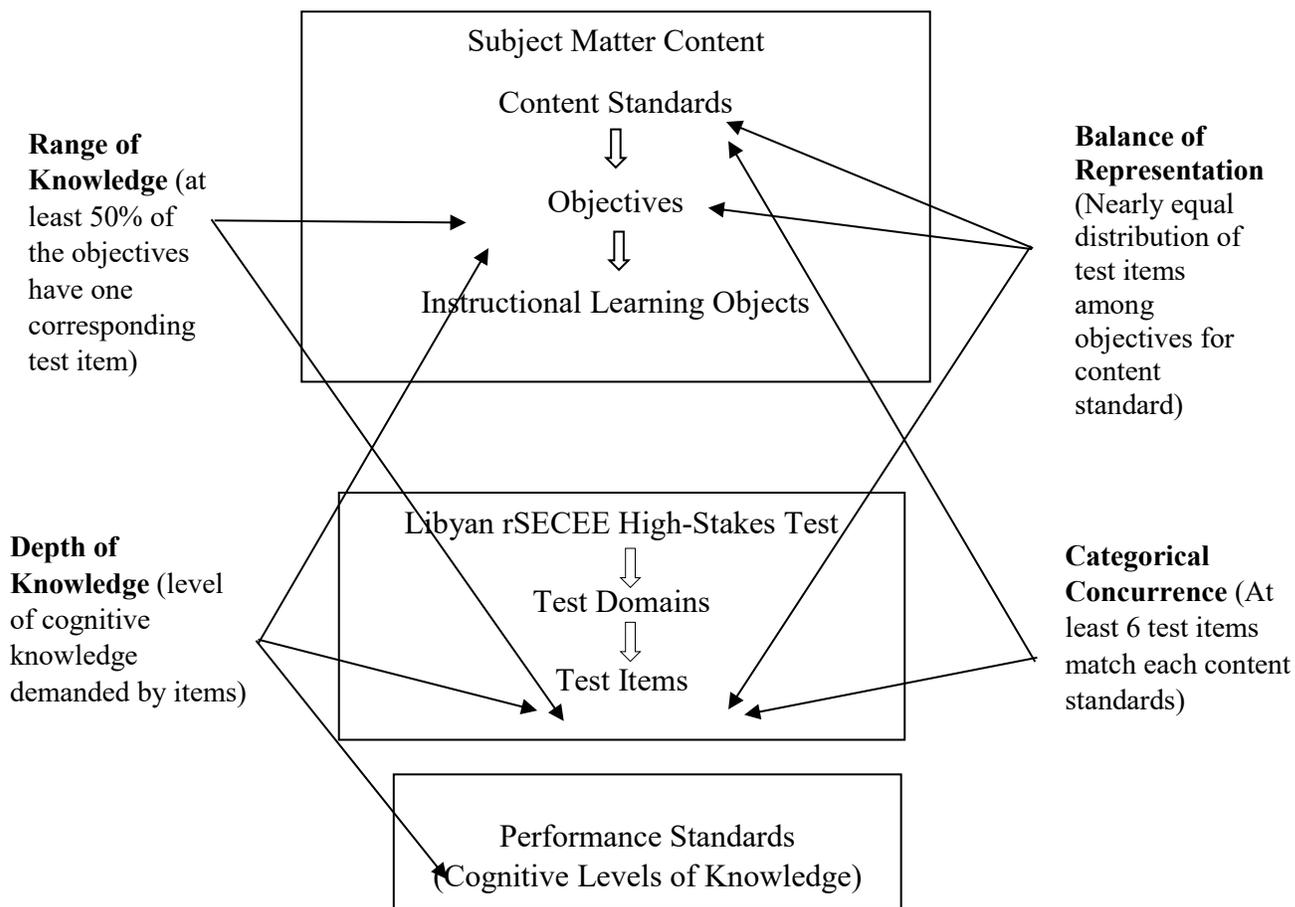


Figure 6.2: Illustration of How Webb's Alignment Criteria Measure the Correspondence between Libya's EFL Content Standards and rSECEE  
Source: adapted from Roach et al (2005).

A follow-up discussion was conducted with the alignment review panel as this provided information about the members' perceptions as to the level of understanding required by Libya's EFL content standards and rSECEE. The discussion also provided an understanding of how the members reacted to the overall alignment process and provided other relevant suggestions.

### 6.3. Analysis

#### 6.3.2. Document Analysis

The researcher employed an inductive analysis approach (Thomas, 2006) for analysing the Grade 12 textbooks. The inductive analysis approach mainly uses “detailed readings of raw data to derive concepts, themes, or a model through interpretations made from the raw data by an evaluator or researcher” (Thomas, 2006, p. 238). In addition, the inductive approach allowed “research findings to emerge from the frequent, dominant, or significant themes inherent in raw data, without the restraints imposed by structured methodologies” (Thomas, 2006, p. 238). The rSECEE for the academic year 2014/2015 also underwent analysis. Importantly, the quality test items’ analysis was not a comprehensive item-by-item analysis where Libyan students’ responses to the 60 test items were examined. It did not consider item difficulty,<sup>23</sup> item discrimination,<sup>24</sup> or reliability from a psychometric perspective. Instead, the analysis was focused on face validity, which is a more superficial evaluation of whether the testing instrument under discussion was a valid measure of the given construct. Face validity differs from other types of validity as it is the “appropriateness, sensibility, or relevance of the test” (Holden, 2010, p.1).

Importantly, two major threats to test validity were considered also in the analysis of rSECEE: “construct under-representation”; and “construct-irrelevant variance” (Messick, 1996).

If a test has the characteristic of construct under-representation it means that the test items which are measured in the test fail to take into account important dimensions of the test construct (Brualdi, 1999). Consequently, the test results may not indicate students’ real abilities within the intended construct. Construct-irrelevant variance means that the test “measures too many

---

<sup>23</sup> A measurement that determines the difficulty level of test items.

<sup>24</sup> A measurement of how well an assessment differentiates between high and low scorers.

variables, many of which are irrelevant to the interpreted construct” (Brualdi, 1999, p.4). In this dissertation, the term *construct* is defined as the subject matter, or domain, being assessed, and learning “implies improved proficiency in the construct variable” (Baird, Andrich, Hopfenbeck & Stobart, 2017, p.318).

### **6.3.3. Alignment Analysis Process**

The analysis, which follows the Webb model (1999), sought to determine the degree of alignment between Libya’s EFL standards and the rSECEE. A minimum level of rater agreement (at least 65% of reviewers agreeing about an item that corresponded to an objective, or the DOK level of an item) was necessary for an item to be considered for analysis or used when reporting results about the degree of alignment between the Libyan EFL standards and the rSECEE. The rationale for such an approach echoes the measurement and communication issues highlighted by Herman et al. (2007). They emphasize that rater agreement serves two purposes: technical and sociopolitical. Herman and colleagues note that rater agreement from a psychometric lens is crucial for reliable measurement within alignment research. From a sociopolitical lens, agreement denotes the degree to which mutual understandings are shared. Therefore, if standards and assessments aim to communicate a common set of expectations; and in accordance with Herman et al. (2007), it would seem reasonable to employ a method that takes into account a rater agreement threshold in judging an item’s attribute (such as the DOK level).

Analysis of the panel members’ responses provided information on the rSECEE attainment based on the Webb (1997, 1999) alignment criteria that were fully discussed in Chapter Three: (a) categorical concurrence; (b) range of knowledge (ROK); (c) balance of representation, and (d) DOK. The first three criteria measure the correspondence between skills and concepts covered in the Libyan EFL content standards, objectives and the skills and concepts tested by the

rSECEE. The last criterion measures the degree of alignment between the cognitive demands of both the Libyan EFL content standards and the rSECEE test items. To ensure an appropriate interpretation of the results based on the Webb model Webb (1997, 1999), the alignment criteria are restated in the following subsections.

#### **6.3.3.1. Categorical Concurrence**

When assessing alignment between standards and assessment, it is fundamental to ascertain whether both components have addressed the same content categories (Webb, 1999). As discussed in Chapter Three, in general terms, the categorical concurrence criterion provides an overall indication of whether the assessed educational components incorporate the same content (Webb, Horton, & O'Neal, 2002).

In this dissertation, six items were set as the minimum requirement for making decisions about students' knowledge and their attainment of a standard. It is worth noting that this number was chosen on the basis that it could yield a "relatively reliable scale for estimating students' mastery of content on that scale" (Webb, 1999, p.7). However, if four to five items were linked to a standard, categorical concurrence was considered weak, and if three or less items correspond to a standard, categorical concurrence was judged unacceptable. In other words, the Libyan EFL content standards and rSECEE were judged to be aligned if an acceptable level was reached for all of the standards i.e., 100%. If 70% or more acceptable level was attained it was considered to be highly aligned, and if there was a 50% to 69% acceptable level of attainment it was considered as partially aligned, and hence, anything less than 50% attainment level was considered poorly aligned. The term *hit* is used in this dissertation to indicate a content standard that has been aligned to an test item.

### **6.3.3.2. Depth-of-Knowledge (DOK) Consistency**

For standards and assessments to be judged as aligned, the complexity of knowledge that each requires must be analysed. The criteria that measures the DOK consistency indicates to what extent the test is cognitively demanding in terms of what the Libyan EFL content standards expect the students to know and do (Martone & Sireci, 2009; Roach et al., 2008; Webb, 1999, 2002, 2007). For this criterion to achieve an acceptable level of alignment, at least 50% of the test items that correspond to an objective had to be at or above the level of knowledge of the actual objective (Lane 2004; Martone & Sireci, 2009; Webb, 1999, 2002). However, the DOK consistency would be judged weak if 40% to 49% of the items corresponding to an objective are written at or above the DOK level of the objective, and unacceptable if 39% or less of the test items were written at or above DOK level of the actual objective. The 50% decision was based on the assumption that most cut-off scores require students to achieve this level in order to pass an assessment (Lane, 2004; Martone & Sireci, 2009; Webb, 1999, 2002). In essence, if at least 50% of the items have to be at or above the DOK level of the corresponding objective then the students would have to answer at least one of these items correctly (Webb, 1999).

Notably, some of the 2014/2015 rSECEE test items had particular characteristics that were difficult to judge, especially due to panel review members' absence of knowledge about instruction. A test item may look challenging and cognitively demanding to a novice but can exemplify a low DOK level, because the knowledge essential to solve it was frequently taught and therefore the students would have had the opportunity to routinely practice similar items during classroom instruction. Therefore, any reading comprehension or grammatical item that was considered a normal recall of information was coded as a DOK level (Level 1), i.e., lower-

order thinking skills because students would have come across, or routinely practiced similar items, within the classroom context.

### **6.3.3.3. Range-of-Knowledge Correspondence**

The ROK criterion compares the breadth of knowledge coverage of standards and assessment. This criterion is employed in alignment research to “judge whether a comparable span of knowledge expected of students by a standard is the same as, or corresponds to, the span of knowledge that students need in order to correctly answer the assessment items/activities” (Webb, 1999, p.8). The ROK criterion examines the distribution of hits across standards. In an ideal scenario, students’ knowledge of a standard would be measured by creating a test that randomly selects items measuring the full domain of content knowledge related to a standard (Webb, 1999). Essentially, if a random sample of items were selected, they would be measuring some knowledge from the full range of objectives.

For alignment on this criterion to be considered acceptable, at least 50% of the Grade 12 Libyan EFL objectives for a standard must have at least one exam item representing it. However, if less than 50% of the objectives under a standard correspond to rSECEE test items, then the ROK criterion is considered unacceptable. The logic behind Webb’s (1999, 2002) assigned criteria is that students ought to be tested on at least one half of the curricular content. Following Webb (1999, 2002), all the Grade 12 Libyan EFL objectives within the ROK consistency category were assumed to have equal weighting and cover the necessary skills to achieve the standards.

### **6.3.3.4. The Balance of Representation**

This criterion judges the extent to which test items are evenly distributed across standards. The index is computed by considering the difference in the number of objectives and the number

of hits appointed to the objective (Webb, 1999, 2002). This index is computed by the following formula:

$$\text{Balance} = 1 - \left( \sum_{i=0}^K \left| \frac{1}{O} - \frac{I_k}{H} \right| \right) / 2$$

Where  $O$  is the total number of objectives hit (i.e., item has been judged to be aligned) for the standard,  $I_k$  is the number of items hit corresponding to objective  $k$ , and  $H$  is the total number of items hit for the standard.

Only objectives for a standard that have one or more hits, in other words a test item corresponding to an objective, are considered for this particular alignment criterion. Therefore, objectives that do not have matching items within the rSECEE were not considered in the equation of balance. As noted by Webb (1999), the balance index can range from zero to one, where zero indicates an unbalanced representation, and one indicates a balanced representation and occurs when hits associated with a standard are evenly distributed amongst the objectives for the given standard. In other words, if the index is close to one, then most of the content objectives have been measured by the same number of test items, and, thus, the test is balanced. In contrast, if the index is close to zero, then only a few objectives are being measured within the test, or the distribution of test items has focussed on only one or two objectives (Bhola et al., 2003). To help clarify this, if 15 objectives for a standard are hit and there were 30 hits, to achieve a balanced representation (i.e., an index value of one) there should be two hits for each objective. Webb (1999) further notes that an index value less than 0.5 represents a unimodal<sup>25</sup> distribution of test items across standards, and an index value of 0.55 or 0.6 represents a

---

<sup>25</sup> Unimodal means there is only a single heights value, i.e. one high peak.

bimodal<sup>26</sup> distribution. An index value of 0.7 or higher indicates that test items are more evenly distributed among the objectives. Therefore, and in accordance with Webb (1999, 2002), the acceptable cut-off point for this criterion is 0.7.

In addition, upon evaluating if a satisfactory level was achieved on a standard for each of the four criteria, no distinction was made if one test item corresponded to more than one objective (i.e., had multiple hits). This decision was based on Webb's (1999) recommendation to help improve the possibility that the test and standards would satisfy the requirements for an acceptable level on two criteria: categorical concurrence and ROK correspondence. It is also worth noting that instead of 60 rSECEE test items, only 58 were included in the following results and discussion. Test items number 28 and 39 were eliminated from the analysis because all reviewers agreed that these two test items were measuring other content (such as prepositions) than the set Grade 12 content objective. The results of the document analysis, and the degree of alignment between rSECEE and Libya's EFL content standards are reported in the following section.

## 6.4. Results

### 6.4.2. English-Secondary EFL Textbooks

Textbooks are the main source of information and a fundamental component of the Libyan educational culture (Abdulhamid, 2011). Within the Libyan context, "textbooks represent the syllabus and dictate what should be taught in the classrooms," and "[t]eachers teach according to textbooks...and achievement tests are designed based on the content of textbook" (Wang, 2006, p.50). Indeed, it can be argued that within the Libyan educational system "textbooks used in classrooms *are* the curriculum" (Richards, 1998, p.125). Students are given textbooks for all

---

<sup>26</sup>A bimodal distribution occurs when data that has two peaks (modes) that are far apart.

subjects and are expected to both comprehend and memorise what is articulated in them in order to pass tests (Abdulhamid, 2011).

The Grade 12 EFL textbooks are designed for students who have successfully completed seven years of EFL schooling with the Libyan education system. The administered textbooks are designed to cover an academic year (from September to June). Adrian-Vallence, Gough and Liz (2014, p.6), who are the authors of the Grade 12 EFL textbooks, state that the overall aims of the textbooks are to:

1. Consolidate and further develop understanding of the grammatical system;
2. Increase the students' range of active vocabulary; and
3. Extend students' ability in the four language skills of reading, listening, speaking and writing.

The revised Grade 12 EFL students' course textbooks for both the literary and scientific sections consist of eight units with each unit having 12 lessons that cover a particular theme for all the language skills such as reading, speaking, writing, vocabulary and grammar. The students' Grade 12 EFL workbook provides ample activities for students to further develop and practice the grammar and vocabulary specified for each unit. For example, Unit 7 covers the theme of health and first aid. The target vocabulary (such as balanced diet, carbohydrate, chest pains and vaccination) is related directly to the theme and students learn new vocabulary and expressions connected to the World Health Organisation, first aid tips and procedures, and safety precautions. In addition, the designated grammar components within the student's course book are built around the theme of each unit.

A summary of how the 12 lessons are distributed among the four language skills and specialised domain of knowledge is summarised in Table 6.1.

Table 6.1.

*Table 6.1: Chronological Distribution of Lessons within the Revised Grade 12 EFL Literary and Scientific Course Textbooks*

| <b>Target Language Skill Within a Unit</b> | <b>Number of Lessons Devoted to the Language Skill</b> |
|--|--|
| Reading                                    | 2  |
| Vocabulary                                 | 1  |
| Grammar                                    | 2  |
| Speaking                                   | 1  |
| Writing                                    | 1  |
| Listening                                  | 1  |
| Specialisation                             | 4  |

Note: The theme of the unit is developed in different directions in line with either the Literary or Scientific specialisation.

Every unit of the revised Grade 12 EFL literary and scientific students' course textbooks starts with a reading component in the target language. The students are presented with ice-breaking tasks that activate and build on prior linguistic knowledge. This is then followed by either a fiction or non-fiction reading text (such as stories, letters, faxes, advertisements, brochures, and newspaper and magazine articles), and reading tasks that aim to promote reading sub-skills including prediction, inferencing reading for specific information and gist. Adrian-Vallence et al. (2014) argue that students are presented with authentic reading texts to enable them "to learn how to deal with a variety of different examples of written English" (p.7). Through the receptive skill of reading and its related tasks, students are introduced to the grammar functions and vocabulary of the unit. These are later developed and consolidated through the productive skills of speaking and writing. In essence, the Grade 12 EFL student course book's philosophy of learning adheres to input and output processing of language acquisition to effectively promote interlanguage development (Clark & Hecht, 1983; Izumi,

2003). Adrian-Vallence et al. (2014) argue that students are not overtly presented with vocabulary and grammar components at the start of unit to enable them to recognise, analyse and understand the target language in use within different situations. Subsequently, these target language components can be enhanced and refined by the “accurate use of the target language in productive situations” (Adrian-Vallence et al., 2014, p.6).

What the Grade 12 Libyan students have acquired from the target language components are put to test towards the end of the unit when the speaking and writing tasks come into play. Adrian-Vallence et al. (2014) emphasize the importance of providing students with ample opportunities for peer feedback and interaction during each unit. Most of the activities in the students’ course book are designed to be answered and checked in pairs. The paired-work tasks, such as role plays, discussions, quizzes, information gap and problem-solving activities, encourage use of prior knowledge and prompt the use of recently acquired vocabulary and expressions from the units.

In addition, speaking tasks such as role plays, discussions, quizzes, information gap and problem-solving activities can be very motivating, because they allow learners to “actively participate in the lesson and to interact successfully in the target language at an early point in the learning process” (Becker & Roos, 2016, p.10). Therefore, the focus of the textbooks is not the form of the language, or the ability to produce a grammatical response. Instead, the focus is on the meaning or general communicative ability of the student, and thus, it can be argued that the textbooks are aligned to mandated Libyan EFL content standards.

Printer (2007) further emphasizes that information gap activities offer language learners the opportunity to “express their own meanings in a less restricted manner” (p.189). Such interaction in the target language would provide opportunities for Libyan Grade 12 students to creatively

experiment with the language, and thus facilitate their acquisition process. Moreover, the writing tasks that are incorporated at the end of each unit encourage a guided writing approach. It is argued that guided writing has the advantage of facilitating language learners' writing process, as well as improving their writing performance (Lan, Hung, & Hsu, 2011; Lee, 1994). Furthermore, the guided writing process can lower students' anxiety level when they are attempting to express themselves in the target language, and enhance students' positive writing attitudes (Lan, Hung, & Hsu, 2011; Lee, 1994). Against this background, Adrian-Vallence et al. (2014) argue that Grade 12 students would be able to produce written assignments (such as a letter) with a more communicative purpose and practical use.

From analysing the speaking and writing tasks (i.e. productive skills tasks) implemented within eight units of the student's Grade 12 EFL course book, it can be argued that these tasks are communicatively oriented and are designed to support students' interaction with their peers allowing them to make use of their "rich resources of imagination, creativity, curiosity, and playfulness" (Zafeiriadou, 2009, p.6, and see Section 2.4). This finding was confirmed by the alignment reviewers during the follow-up discussion about the rSECEE. They all agreed with Reviewer A's statement:

Reviewer A: We had an opportunity to look at the textbooks, it's a commutative text book, with its thematic based, EAP, commutative language teaching.

Moreover, it was evident from the Grade 12 EFL students' textbooks (course and work books) that the revised curriculum has been designed to provide students with a wide range of tasks to promote language learning within a meaningful context, and which offers students diverse opportunities to become autonomous language users. The materials presented within the textbooks can be described as somewhat relevant to both the scientific and literacy streams. Like many curricula revisions, the revised secondary level EFL curriculum, learning material, and

methods of delivery are designed to promote students' language learning, collaboration and motivation. The students' course book is full of challenging, yet stimulating, authentic tasks that encourage students' engagement. The incorporated tasks methods build on "collaborative learning among students as they search for information and test their emerging knowledge with other people" (Paris, 2000, p.6). These authentic tasks have replaced the traditional workbook exercise and monotonous drilling. Educational research, such as Turner and Paris (1995), has demonstrated that tasks necessitating student interaction and involvement promote better opportunities for learning as students find these tasks more engaging and challenging than tasks that require students just to complete a task. A sample unit of the Grade 12 EFL literacy course textbook is found in Appendix O.

It is worth noting that alignment review panel members' reflections and comments on the Grade 12 EFL students' textbooks during the follow-up discussion interview were in line with the reported results. The following quotation clearly illustrates this finding.

Reviewer A: I would say that those are not bad textbooks, in the hands of a creative teacher, I think those potential learning outcomes could be realised with the information there, I think if you are going to have textbook in a country, it's not a bad one to have.

As Phase I seeks to answer the first question, "To what degree is the rSECEE aligned with Libya's EFL standards?", it was fundamental in the document analysis to examine if the Libyan content standards were operationalised in the Grade 12 student course book. From the analysis, it can be safely stated that the six EFL content standards were operationalised through the course book, and the textbook tasks necessitated a range of DOK levels for the Libyan Grade 12 students. Table 6.2 illustrates examples of how the six Libyan EFL standards are operationalised within the assigned textbooks.

The findings were confirmed and validated by the alignment review panel members who agreed with Reviewer A's comment. Reviewer A reported that the EFL content standards were operationalised to a certain degree within the textbooks, there were a variety of tasks operationalising these standards, and the four language skills were represented to a great degree within the textbooks.

Reviewer A: I could see a great variety in terms of what those [EFL content standards] you know whether it was a content analysis or skills we saw a variety in the textbook in relation to the outcomes. And we saw a range of potentialities with these learning outcomes as the ministry goals and emmh in terms of the textbook... to a degree there is a potential for all those outcomes to be operationalised by the textbook, provided that the teacher translates them in that way...the four skills are represented ...the reading, speaking listening and writing extensively.

Reviewer B: we all agree...we felt that there is potential ...there is a variety of tasks.

Reviewer C: its there (refereeing to standards in the textbook).

Furthermore, the reviewers agreed that a range of DOK levels were being operationalised within the Grade 12 EFL textbooks, and the levels of thinking necessitated on the part of the Libyan students extended beyond simple recall of information.

Reviewer A: There is a quite a range and that goes beyond, we go beyond, simple recall of information.

Table 6.2

*Examples of the Libyan EFL Standard Being Operationalised within the Assigned Textbooks*

| Student's course and work book tasks  | Target Skill  | Curricula Objective  | Operationalised Standard |
|---|---|--|--------------------------|
| <p><b>Discuss these questions in pairs.</b></p> <ol style="list-style-type: none"> <li>Are texts 1 and 2 good beginnings of a novel? Why?/Why not?</li> <li>Which one makes you want to continue reading the most?</li> <li>How do you decide whether to start reading a book? Do you read any parts of the book to help you decide?</li> <li>How does the type of narrator – the third person in text 1 (<i>He raised himself...</i>) and the first person in text 2 (... <i>I can't be sure.</i>) – change the style of the text?</li> </ol>  | <p>Reading<br/>(Unit 5<br/>work book:<br/>p.32)</p>                     | <p>R. 22: To<br/>practice<br/>reading fiction.</p>   | <p>Standard 5</p>        |
| <p><b>B Work in pairs. Explain the meaning of each sentence in your own words.</b></p> <p><b>Example:</b> He'll be in Cairo until 8 p.m. <u>At 8 p.m., he will leave Cairo.</u></p> <ol style="list-style-type: none"> <li> <ol style="list-style-type: none"> <li>He'll be in Cairo until 8 p.m. _____</li> <li>He'll be in Cairo by 8 p.m. _____</li> </ol> </li> <li> <ol style="list-style-type: none"> <li>I won't finish until lunchtime. _____</li> <li>I won't have finished by lunchtime. _____</li> </ol> </li> <li> <ol style="list-style-type: none"> <li>I'll work until the programme starts. _____</li> </ol> </li> </ol>  | <p>Vocabulary:<br/>(Unit 3,<br/>lesson 3,<br/>p.32)</p>                 | <p>V. 4: To<br/>practise using<br/><i>by, until</i> and<br/>other words<br/>and phrases<br/>used to refer to<br/>the future.</p> | <p>Standard 1</p>        |
| <p><b>B Some of the sentences are wrong. Tick (✓) the ones that are right. Correct the ones that are wrong.</b></p> <ol style="list-style-type: none"> <li>If the patient feels sick, he should be asked to sit down.<br/>_____</li> <li>The new stadium will finish in three years.<br/>_____</li> <li>This medicine can take by adults and children over the age of eight.<br/>_____</li> <li>You won't be allowed to leave the room until the exam is over.<br/>_____</li> <li>The dish can be make with fresh or dried pasta.<br/>_____</li> <li>People who are suffering from shock shouldn't be left alone.<br/>_____</li> <li>The prize might give to my uncle this year.<br/>_____</li> </ol> | <p>Grammar:<br/>Unit 7,<br/>lesson 4<br/>task B work<br/>book, p.44</p> | <p>G. 24: To<br/>recognise<br/>written<br/>grammar<br/>mistakes</p>  | <p>Standard 2</p>        |

|  |  |   |                   |
|--|--|---|-------------------|
| <p><b>C</b> Now ask your partner the questions you wrote in Exercise B. Then answer your partner's questions. How many differences can you find between the two texts? Note: Do not read your text aloud, and do not read your partner's text.</p>   | <p>Speaking:<br/>Unit 6,<br/>lesson 6,<br/>p.71.</p>   | <p>S.18: To exchange information to find inconsistencies</p>        | <p>Standard 4</p> |
| <p>Use your notes to write a paragraph about the book in your notebook. Give information about the book, briefly tell the story and give your opinion. Use phrases from lesson 6.</p>  | <p>Writing:<br/>Unit 5,<br/>lesson 7,<br/>p.60.</p>    | <p>W.16: To write a book review.</p>                                | <p>Standard 6</p> |
| <p><b>L</b> Listen to part 2 again and complete the instructions.</p> <ol style="list-style-type: none"> <li>1. Hold him _____ and slap his back.</li> <li>2. Do it quite hard, but _____ not to hurt him.</li> <li>3. The fingers should meet just above his _____.</li> <li>4. One or two quick presses should clear the _____.</li> <li>5. Make sure you clear his _____ afterwards.</li> <li>6. Never give water to someone who is _____.</li> </ol> | <p>Listening:<br/>Unit 7,<br/>lesson 12,<br/>p.89.</p> | <p>L.23: To practice listening to information and instructions.</p> | <p>Standard 3</p> |

Note: R, V, G, S, W, and L stand for Reading, Vocabulary, Grammar, Speaking, Writing and listening respectively.

However, the reviewers argued that the DOK levels were not evenly distributed within the textbooks. They stated that the majority of content objectives operating within the eight units of the student course book were uniform in nature as they required a DOK level of no more than Level 1 and Level 2, but then a DOK Level 4 would appear. One reviewer illustrated that the content objectives designated for grammar mostly targeted DOK Levels 1 and 2, but there were objectives that necessitated a dramatic change in the student's thinking level to Level 4:

Reviewer A: The DOK levels are scattered throughout [the listed objectives], you would go from a 1 and 2 [DOK] reoccurring throughout the unit and all of a sudden you would pop up to 4... in the grammar component you go from practicing the verb to be up to identifying styles of writing and writing an article about a mysterious place ...really there's a jump.

Rationalising the prevalence of Level 1 and 2 content objectives in the textbook the reviewers stated that these Levels are foundational for reinforcing learning in any learning process:

Reviewer A: They are a lot of ones and twos, but when you go into the writing you get more fours and threes... it would make sense for them to have more ones and twos throughout because they are the foundations for reinforcing learning.

The following section presents the document analysis results for the 2014/2015 rSECEE's.

### **6.4.3. The rSECEE**

Similar to the other rSECEEs, the rSECEE automated scoring for 2014/2015 academic year follows a discrete-point testing format with 60 items: 25 true/false items; 25 MCQs; and 10 matching items. The True-False items present the test-takers with a “statement and ask the respondent to discern whether the statement is true” (Stone, 2001, p.11). Multiple-choice items present the examinees with either a question or incomplete statement and three to five possible answers including distractors<sup>27</sup> (Stone, 2011). Most MCQs had distracters similar in length, complexity and grammatical form to the correct answer. Although three option answers can be adequate, four options were employed in all the MCQs. The four-option approach, which is termed the “Goldilocks formula” may have been employed in the rSECEE to help maintain the validity of a question stem and overall test (Zarza & Abedalazeez, 2014). The options, as noted by Hubley (2012), represent the following:

- a. One option – the key, at just the right level of specificity.
- b. One option that is too general for a main idea.
- c. One option that is too narrow, often citing a supporting detail or example.

---

<sup>27</sup> They contain several incorrect answers or distractors. “They are named distractors because their plausibility should draw away test-takers who are not secure in their knowledge base. Distractors are listed along with the correct answer” (Stone, 2001, p.12).

- d. One option that is related, but off topic.

Matching items normally present a list of terms/words and a list of definitions and the examinee is asked to match the item with the definition (Stone, 2001). The overall formatting of the item stems<sup>28</sup> were a partial sentence rather than question stem format.

Similar to Onaiba's (2014) findings, it was evident from the quality test items analysis that the 2014/2015 rSECEE made no effort to measure Libyan students' EFL language competence for the reading, listening, speaking, and writing skills. Instead, the 60 test items appear to be testing basic recall of grammar rules, vocabulary knowledge, and content knowledge articulated within the Grade 12 EFL student textbooks. Approximately 30% of the test items measured simple recall of grammar knowledge or rules, and 23% and 47% of the test items measured the examinees' knowledge of vocabulary and Grade 12 EFL student course textbook content information respectively. Thus, it can be safely claimed that the 2014/2015 rSECEE has not shifted the focus from linguistic knowledge to language use. The ultimate focus of the secondary level examination is still on the linguistic knowledge of English.

The findings from the document analysis of the Libyan EFL content standards for Grade 12 content objectives and textbooks indicate that the rSECEE may not be accurately measuring the target construct. Although the Libyan content standards have been operationalised to a degree through the Grade 12 EFL content objectives being articulated within Grade 12 EFL textbooks (as reported in Section 6.4.2), the opposite is found when the 2014/2015 rSECEE is evaluated. A limited number of content objectives are measured that involve a lower level of thinking by the students. The above face validity findings were substantiated through Webb's (1997, 1999) comprehensive validation tool (see Section 6.4.4.).

---

<sup>28</sup> The problem or central question presented in the item.

Furthermore, many of the 2014/2015 test items have been poorly crafted. The high-stakes test consists of a number of ambiguously worded test items, questions written in non-standard forms, item questions testing too much information and not reflecting a specific content, questions without a stem, as well as spelling errors. Table 6.3 summarises examples of poorly crafted test items in the 2014/2015 rSECEE.

Table 6.3

*Examples of Poorly Crafted Test Items in the 2014/2015 rSECEE*

| Test Item Number | Item   | Item Problem  |
|------------------|--|---|
| 9                | He said he had done it, yesterday.           | Ambiguity, not reflecting a specific content, and a question without a stem.    |
| 13               | They would have won if they had worked hard. | Ambiguity and not reflecting a specific content.                                |
| 16               | I'll be taking my final exam next Sunday.    | Ambiguity and not reflecting a specific content, and a question without a stem. |
| 17               | We expected him coming soon.                 | Ambiguity, not reflecting a specific content, and a question without a stem     |
| 24               | My uncle is kind for me.                     | Ambiguity, not reflecting a specific content, and a question without a stem.    |
| 35               | The Red Crscent was set up ...years ago.     | Spelling errors   |
| 47               | “bored” is used to say...people feel.        | Non-standards form of punctuation.  |

The ambiguity in the True-False items listed in Table 6.3 (i.e. test items 16, 17 and 24) may cause anxiety and confusion on part of the test-takers, as it is not clear what the item is testing. For example, test item number 24 “*my uncle is kind for me*” is grammatically incorrect, although knowledge-wise it may be correct. In essence, if the students think that the item is measuring their grammar competence then they will answer false, but if they consider it as piece of knowledge relating to one’s personal experience they may answer true. Thus, the students’ answers are jeopardised. Poorly crafted test items may cause a competent student to answer incorrectly, or a student who has no knowledge of the material to answer the test item correctly.

Therefore, the item flaws may constitute construct irrelevant variance (CIV) which can incorrectly inflate test scores (Downing, 2002). The construct underrepresentation (CUR) and CIV in relation to 2014/2015 rSECEE are discussed further in Section 6.5.

#### **6.4.4. Webb Model (1997, 1999) Alignment Results**

Overall, the alignment review panel members achieved a high degree of agreement in judging the alignment between Libyan EFL content standards and the rSECEE (see Table 6.4).

Table 6.4

##### *Reviewers' Agreement on Coding*

| <b>Reviewers (N)</b> | <b>Reviewers' Agreement on Coded Standards</b> | <b>Reviewers' Agreement on Coded Objective</b> |
|----------------------|--|--|
| 4                    | 0.83   | 0.73   |

##### **6.4.4.1. Categorical Concurrence**

The categorical concurrence criterion is met if similar content is addressed in both standards and assessment. This criterion was assessed by establishing if the rSECEE included items measuring content from each Libyan EFL content standard. The data in Table 6.5 indicates that the categorical concurrence for the rSECEE is unacceptable, as only two of the six set standards were at the acceptable level.

Table 6.5

*Categorical Concurrence of the rSECEE*

| Libyan EFL Standards | Number of Objectives | Number of Hits | Mean Number of Hits | Categorical Concurrence Level | Percent Acceptable |
|----------------------|----------------------|----------------|---------------------|-------------------------------|--------------------|
| Standard 1           | 61                   | 20             | 0.35                | Acceptable                    | 33.3%              |
| Standard 2           | 23                   | 38             | 0.65                | Acceptable                    |                    |
| Standard 3           | 40                   | 0              | 0                   | Unacceptable                  |                    |
| Standard 4           | 5                    | 0              | 0                   | Unacceptable                  |                    |
| Standard 5           | 2                    | 0              | 0                   | Unacceptable                  |                    |
| Standard 6           | 14                   | 0              | 0                   | Unacceptable                  |                    |
| <b>Total</b>         | 145                  | 58             |                     |                               |                    |

In summary, the data indicates the attainment level of categorical concurrence criteria was below 50%. Therefore, the rSECEE can be deemed to be poorly aligned and did not meet Webb's (1999, 2002) alignment criteria for categorical concurrence.

**6.4.4.2. Depth-of-Knowledge (DOK) Consistency**

The DOK percentage of items for the rSECEE compared to the target DOK (the actual DOK level of the tested objectives) are presented in Table 6.6. There was a significant difference between the DOK levels from the 2014/2015 rSECEE and the target distributions.

Table 6.6

*A Comparison of DOK Target Levels to the rSECEE DOK Levels*

|         | Target DOK Level (%) | Actual DOK for the rSECEE (%) |
|---------|----------------------|-------------------------------|
| Level 1 | 62.10                | 100                           |
| Level 2 | 24.14                | 0                             |
| Level 3 | 10.34                | 0                             |
| Level 4 | 3.43                 | 0                             |

After comparing each item and its assigned objective, and indicating the DOK level was either below, at or above the actual DOK Level, the DOK consistency for the rSECEE was examined by standard (see Table 6.7).

Table 6.7

*DOK Consistency for the rSECEE*

| <b>DOK Consistency</b> | <b># Items Below the DOK Level</b> | <b># Items at the DOK Level</b> | <b># Items Above the DOK Level</b> | <b>DOK Level</b> | <b>% at or above the DOK Level</b> |
|------------------------|------------------------------------|---------------------------------|------------------------------------|------------------|------------------------------------|
| <b>Standard 1</b>      | 8                                  | 12                              | 0                                  | Acceptable       | 60.0%                              |
| <b>Standard 2</b>      | 14                                 | 24                              | 0                                  | Acceptable       | 63.2%                              |
| <b>Standard 3</b>      | 0                                  | 0                               | 0                                  | Unacceptable     | 0%                                 |
| <b>Standard 4</b>      | 0                                  | 0                               | 0                                  | Unacceptable     | 0%                                 |
| <b>Standard 5</b>      | 0                                  | 0                               | 0                                  | Unacceptable     | 0%                                 |
| <b>Standard 6</b>      | 0                                  | 0                               | 0                                  | Unacceptable     | 0%                                 |

An analysis of each DOK level by the measured standards is presented in Figure 6.3. The rSECEE items' DOK were inconsistently distributed. All test items (n = 58) were at the DOK Level 1. The results in Tables 6.6, 6.7 and Figure 6.3 indicate that the rSECEE demonstrates acceptable levels of DOK consistency on only one-third of the Libyan EFL content standards, with only Standards 1 and 2 exhibiting items meeting the DOK levels of the objectives. Thus, the SECEE was successful in achieving an acceptable level on the DOK consistency criterion for only two standards.

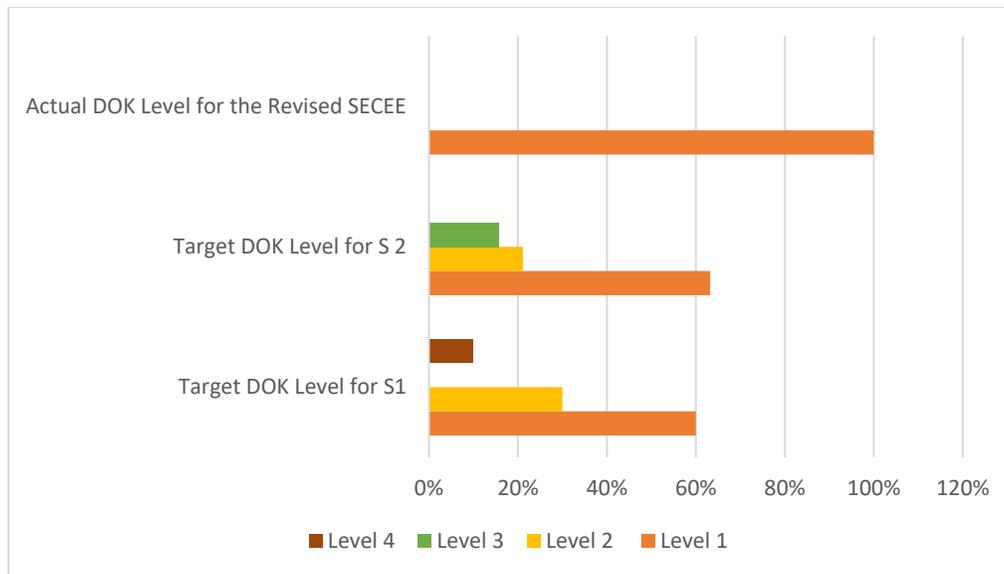


Figure 6.3: Distribution of DOK Level of SECEE Test Items by Standards  
Abbreviations: S 1 and S 2 = Standard 1 and Standard 2, respectively.

#### 6.4.4.3. Range-of-Knowledge Correspondence

Since the rSECEE was implemented, the convention has been to test the students' knowledge with only 60 items. According to Webb (1999, 2002), this restriction on the number of test items may place an upper limit on the number of objectives that can be assessed. Table 6.8 presents the ROK level for the rSECEE.

Table 6.8

#### *Range of Knowledge Criterion for the rSECEE*

| SECEE             | Number of Objectives | Number of Objectives with Hits | ROK Level    | Percentage % of Objectives Hit |
|-------------------|----------------------|--------------------------------|--------------|--------------------------------|
| <b>Standard 1</b> | 61                   | 12                             | Unacceptable | 19.7                           |
| <b>Standard 2</b> | 23                   | 10                             | Unacceptable | 43.5                           |
| <b>Standard 3</b> | 40                   | 0                              | Unacceptable | 0.0                            |
| <b>Standard 4</b> | 5                    | 0                              | Unacceptable | 0.0                            |
| <b>Standard 5</b> | 2                    | 0                              | Unacceptable | 0.0                            |
| <b>Standard 6</b> | 14                   | 0                              | Unacceptable | 0.0                            |
| Total             | 145                  | 22                             |              | 15.2                           |

Overall the 2014/2015 rSECEE only tested 22 objectives of the 145 set objectives, i.e., only 15.2% of the Grade 12 EFL content objectives. Twelve of the tested objectives were measuring Standard 1, and the remaining 10 objectives were measuring Standard 2. As seen in Table 6.8 the rSECEE did not achieve acceptable levels across all six standards for the ROK criterion. Thus, DOK consistency and ROK correspondence were the two criteria that received the lowest number of standards and assessments. In other words, all 2014/2015 rSECEE test items were at a level of knowledge (Level 1 - recall) below that of the objectives the item was to measure and a very small number of the objectives for a standard had related items in the assessment; hence, this finding validates the face validity finding reported in Section 6.4.2

#### **6.4.4.4. The Balance of Representation**

The SECEE Standard 1 has 61 objectives, but only 12 objectives had hits within the 2014/2015 revised SECEE. There were 20 hits distributed across the 12 objectives, with one objective having five hits, eight objectives had one hit each, two objectives had two hits each, and one objective had three hits; the formula for the Standard was calculated as such:  $1 - (|1/12-5/20| + |1/12-1/20| + |1/12-1/20| + |1/12-1/20| + |1/12-1/20| + |1/12-1/20| + |1/12-1/20| + |1/12-1/20| + |1/12-1/20| + |1/12-1/20| + |1/12-2/20| + |1/12-2/20| + |1/12-3/20|)/2$  such that  $1 - 0.69999/2 = 1 - 0.3499 = 0.6501$ .

Standard 2 has 23 objectives, but only 10 objectives within the 2014/2015 rSECEE had a total of 38 hits; one objective had 20 hits, 4 objectives had one hit each, 3 objectives had two hits each, one objective had five hits, and one objective had three hits. The formula for the Standard 2 was calculated as such:  $1 - (|1/10 -20/38| + |1/10-1/38| + |1/10 -1/38| + |1/10-1/38| + |1/10-1/38| +$

$(|1/10-2/38| + |1/10-2/38| + |1/10-2/38| + |1/10 -5/38| + |1/10 -3/38|)/2$  such that  $1 - 0.915789/2 = 1 - 0.4578 = 0.521$ .

For the balance of representation criterion reported in Table 6.9 the six EFL standards did not meet the acceptable level in the rSECEE.

Table 6.9

*Balance of Representation Criterion for the rSECEE*

| The rSECEE | Number of Total Objectives | Number of Objectives Hit | Number of Total Hits | Balance Index | BOR Level    | Percentage |
|------------|----------------------------|--------------------------|----------------------|---------------|--------------|------------|
| Standard 1 | 61                         | 12                       | 20                   | 0.650         | Unacceptable | 0%         |
| Standard 2 | 23                         | 10                       | 38                   | 0.521         | Unacceptable |            |
| Standard 3 | 40                         | 0                        |                      |               | Unacceptable |            |
| Standard 4 | 5                          | 0                        |                      |               | Unacceptable |            |
| Standard 5 | 2                          | 0                        |                      |               | Unacceptable |            |
| Standard 6 | 14                         | 0                        |                      |               | Unacceptable |            |

The following section presents the alignment reviewers' opinions on the quality of the standards and the rSECEE test items during follow up discussion, and the test developer's responses on the test development process. This is then followed by a discussion of the reported results on the degree of alignment between the rSECEE and Libya's EFL content standards.

#### 6.4.4.5. Panel Review Responses

In the follow up discussion, the alignment panel members reported that the learning outcomes are communicative based and necessitate a wide range of learning skills that not only build on and promote language learning skills, but also life-long learning skills such "goal setting and valuing"

Reviewer A: If we look at the learning outcomes, we can see they are communicative. More than just a system of the language, but we see an increased knowledge of the lexical

system, discussing topics related to specializations, autonomy, goal setting, valuing, [and] raising awareness of the world, those, by and large, are the key learning outcomes... if we look at the learning outcomes, you can see there is quite a range in terms of the depth of knowledge, from simple item identification and recall to issues of valuing, critical thinking, creating, [which are] extended thinking from recall.

In addition, all reviewers reported that the Grade 12 EFL construct as set by the Ministry of Education has been grossly underrepresented by the 2014/2015 rSECEE and the content was tested on trivial information to language acquisition and learning. Furthermore, all the 60 test items were at the lowest DOK level (i.e. Level 1) requiring no more than factual recall of information or rules. This validates the findings of the DOK and ROK criterion.

Reviewer A: Recall of narrow content is basically what is being tested. But basically, here just recalling content or information presented in the textbook. You are actually not reading, you are not writing, you are not speaking, and you are not listening, there is no reading ...so its recall of information, memorization. Even though there's nothing to read and its recall, but they are actually recalling a detail, it is testing information that have been taught...they're testing whether they can recall that bronze was made. I think it is more of a detail, reading for detail...Memory and reading, reading for information and study skill guessing... it's hard to imagine that you have a speaking test without testing speaking, its hard to imagine that you have a writing test without testing writing if they don't have a writing component.

Reviewer B: It's the wrong construct... you are basically testing memory and attendance (all the reviewers then laugh) ... I don't see anything that goes beyond remembering the text... you have already got underrepresentation of construct. The construct its representing is memory, vocabulary and bit of grammar.

Reviewer C: I found most of the items to be either of reading, reading for specific information, or grammar point, some quite obscure. The test is decontextualized and testing recall of information.

Reviewer D: Totally all on recall and its based-on memory of vocabulary and bits of grammar...the test is decontextualized, if it's meant to be a gate keeper, and its doing its job.

The reviewers further reported that despite the fact that the Libyan EFL content standards are operationalised within the textbook, there was no evidence of them being operationalised within the 2014/2015 rSECEE, hence echoing the categorical criterion finding reported in Section 6.4.4.1. Reviewer A explained that it is very problematic to see the curriculum poorly operationalised within the rSECEE by presenting English in discreet isolated points (i.e. in a mechanical manner), and thus not a proper representation of authentic language usage.

Reviewer A: What is important about the exam, is that the exam is intended to reflect the learning outcomes in the textbook which intend to reflect the learning outcomes of the ministry, for the test we have a totally different story...and yet the textbook is designed to develop those skills and clearly the learning outcomes are designed or developed with those major aims in mind... the problem is the textbook isn't designed to be used this way [meaning to be tested this way]. It is very clear from the suggested activities exercise etc. in the class we have a gross underrepresentation of the construct. If the construct is operationalised by the textbook, which I think it is, then this test grossly under-represents it and defines English as a mechanical recall of discrete information. You know there is no evidence that you could use English for anything or in any way except memorising.

Reviewer D: The construct here is underrepresented in the test and not all the objectives are operationalised in the test items.

In accordance, with the document analysis of the rSECEE findings, the panel reviewer members emphasized that the 2014/2015 rSECEE had poorly crafted items. They reported that the rSECEE consisted of misleading items that had item questions without a stem, ambiguous items that did not reflect specific content, and items with spelling errors.

Reviewer A: We don't know what they're testing here; test item number 9 for example "He said he had done it yesterday", and that is true for a number of items here. We don't know what they are actually testing...look circle questions 9 and 13... its missing a stem

Reviewer B: It's [referring to item number 20] a reading item but they are testing it as a grammar... I don't say anything that goes beyond remembering the text. Your 9 and 13 are interesting because you don't know quite what you have to do. It's incorrect (said with a high intonation) ...You would say false because it's incorrect.

Reviewer C: Oh, it's awful, but this is very common. It happens everywhere where there is rating where people's lives are secured as a result of marking failure. I think this is the nature of such sort of tests regardless of context.

Reviewer D: the problem is there is no answer to the question. How would that make you feel? It's determining if you go on or not?

Reviewers strongly emphasised that the 2014/2015 rSECEE could be considered as unethical because it presented uncompleted questions<sup>29</sup> having no exact answers to students. In addition, these chosen test items had nothing to do with language learning and encouraged the narrowing of the curriculum and teaching to the test. Thus, there was a possibility having negative washback on the Libyan EFL classroom.

Reviewer A: NO test question should be incomplete, it should be proofread for spelling errors, and it's not Kosher [meaning not fit or right] in any testing context to give a student in a high-stakes test a question where there is no answer. And a couple of times here there is no correct answer because they have incorrectly included a verb.

Reviewer D: But it's not about going to university, so it makes the students want to pass the test and get prepared for the test in the sense 'what do I need to pass the test' 'what strategy do I need to learn so I don't have to worry about anything else'.

Reviewer B: (the reviewer putting herself in the shoes of a Libyan student) I would focus my attention on the textbooks and the meaning, I'd be teacher dependent in the sense that

---

<sup>29</sup> When a question is missing a content that is essential for providing students with a chance of answering the question

I would expect that she might guide me - this is important keep an eye on this and if you are going to get question as bronze was or Nazca lines. I would be paying attention to content. It drives you back to the textbook; that's the whole point of it seems to me.

The reviewers further argued that this high-stakes test may not be accurately representing a test that is supposed to be measuring language proficiency. Instead, it was more or less an achievement test and an evaluation of whether students had attended their EFL classes and were able to memorise and recall information. More interestingly, Reviewer B reported that the test does not discriminate between what students have understood or memorised from the classroom instruction, and that it is operating as a gate keeper for the Grade 12 EFL students.

Reviewer B: [The objective of the test is] to identify those students who have not focussed in class, who have not paid attention and who have not attended, and then to test memory... in that they have to remember. You need the context of the classroom. It's difficult to separate what they have understand and memorised...if the purpose is to have an easy to mark test and to stop people from going where they want to be going, then they have achieved their aim...this is an achievement test.

Reviewer A: Yeah, it's an achievement.

Reviewer C and D: Definitely, we agree.

The following section discusses the research findings to judge to what degree the rSECEE and the Grade 12 Libyan EFL standards are aligned.

#### **6.4.5. Test-Developer's Responses**

In the interview, the test developer reported that he had participated twice in the development of the rSECEE. The test developer further noted that language-test developers in Libya receive no training for test development, statistical methods, and test validation. Recognising the importance training for the test development process, the test developer reported that many test developers and language inspectors have considered accessible training opportunities. For

example, they have asked test experts who have experience in psychometric testing and effective test development strategies to deliver workshops and short training sessions.

Test Developer: I've developed a number of tests and I've participated twice in the development of the rSECEE...As language inspectors we are selected by the Ministry of Education to develop one of the two high-stakes tests...You may be surprised if I tell you that we have had no training whatsoever when it comes to test development. This may sound unrealistic, but they [Ministry of Education] assume that we can develop a good test since we are language teachers and have some idea about the target curriculum. But in reality, we have no knowledge of effective test development strategies and how tests can be validated. As an inspector and test developer, I can totally see the importance of such training and how it's a must for the test development process. Therefore, we've [language inspectors and test developers] asked assessment experts holding doctoral degrees in assessment and evaluation to help us out, and give us workshops about effective test-development strategies and the issues of reliability and validity...I really believe training is very important for us.

It was further reported that test developers of either the rBECEE or rSECEE are not provided with any test specification. The Libyan Ministry of Education provides the selected test-developers with only the fixed test format. In addition, the developed test items are administrated to Grade 12 students in Libya without undergoing any piloting, or item validation.

Test developer: What happens in Libya is very different from the test development procedures that you've just describe. In Libya we are not given any test specifications or any criteria for development. Test developers don't meet together and select items from a pool of items, there is nothing of this. No one will even question the reliability of your items. The only instructions we are given is that we need to develop two tests and they have to readable. By this time all test developers know that the test must have 60 test items: 25 true/false items; 25 MCQs; and 10 matching items. You develop the test and you type it and then enclose it in an envelope and then it's ready to be sent to the Ministry.

The test developer further argued Grade 12 EFL construct as set by the Ministry of Education has been grossly underrepresented, and is an achievement test. It is an evaluation of whether students are able to memorise, recall information, and apply the grammar rules that have been taught. In essence, the rSECEE is not accurately representing a test that is supposed to be measuring language proficiency, hence echoing the panel review responses reported in Section 6.4.4.5.

How can you have a test [language test] and have students passing the test and the students are unable to construct or articulate grammatically correct sentence...The exam does not measure the four language skills especially with the revised examination that is currently in place. For example, how is the exam evaluating writing here, speaking there is no speaking there is only grammar and vocabulary. How are we supposed to measure student's language ability in all language skills like speaking, listening and writing with such examination in place. What we are testing today is just grammar and vocabulary, the test is testing whether the student knows the meaning of the target vocabulary, grammar, the application of grammar rules...This isn't a language proficiency test...it doesn't tell me anything about the student's language proficiency level, this is an achievement test.

Moreover, the test developer argues that the current rSECEE does not measure what it is supposed to measure, which could lead to imprecise inferences being drawn from test scores and providing an inaccurate picture about Libyan students' language ability. The following quotation clearly illustrates this finding.

The high-stakes test that we currently have [i.e. in Libya] does not provide an accurate measurement of students' language ability. For example, if a student got 97% it does not necessarily mean that this student is highly proficient in English. But what this high score represents is the student's ability to apply grammatical rules that he/she has learnt throughout the year. By the way, this is not only my opinion, it is the opinion of many language inspectors and curriculum evaluators. What I've noticed is that many students have passed

the exam and they don't know anything about the language...I have come across students who are actually at the university level now and they have passed these high-stakes test with 95% and are unable to read or explain the meaning of a sentence. They basically just passed the tests by following grammar rules without any knowledge of the language.

The test developer's further reflections and comments on the rSECEE were in line with the reported results in Section 6.4.4. The test developer argued that the components of Libyan education system are not functioning in harmony, and the ultimate focus of the rSECEE is still on the linguistic knowledge of English.

I have come to a conclusion that the main aim of teaching and learning English in Libya for students is to pass the exam and not learn a language. The most important point that I want to emphasize today is that, teaching, assessment and learning are not working towards the same goal which is a student learning a language. Teaching is tailored towards passing the test and that's it and not towards learning the language. The exam is just testing the student's linguistic knowledge and not the communicative knowledge of the language.

## **6.5. Discussion**

In any educational system, including the Libyan context, there are many participants: people who construct the standards, teachers who work with and assess students, and those who are responsible for constructing high-stakes assessments. Because of this huge number of participants, there may be a risk that the components are not functioning in harmony (Näsström, 2008).

An approach recommended by Webb (1997) was employed in this dissertation to examine the degree alignment of a high stakes test to Libya's EFL content standards. The approach combined both qualitative expert judgments and quantified coding to evaluate the alignment of standards and assessments (Flowers et al., 2006). The product of the analysis was a set of statistics illustrating the degree of alignment between the content knowledge articulated in the

content standards and the content knowledge in the target assessment. Four criteria were used to judge the degree of alignment: categorical concurrence, DOK consistency, ROK correspondence, and balance of representation. Importantly, in any study such as this one, employing the Webb (1997) approach demonstrated the “relationship between what is being asked of the students, how that is being assessed, and what trade-offs are made in the process” (Martone & Sireci, 2009, p.1342). In addition, I am not looking for a *tight alignment* between the rSECEE and the Libyan EFL content standards. Tight alignment necessitates that content standards provide a comprehensive description of the content objective, and that each objective is accurately measured (Linn 2005). However, as advised by Linn (2005), a tight alignment or total congruence is unrealistic; the “goal in studies of alignment is to evaluate the degree to which it is approximated” rather than fully achieved (p.6). Furthermore, according to Looney (2009), tight alignment has the tendency to weaken the quality of innovative programmes where the learning objectives can exceed the ones highlighted in the standards and prioritised in high-stakes tests. Therefore, an “acceptable degree of alignment” is the outcome I sought in this dissertation research.

Similar to Webb et al. (2002), a number of findings implicit in the relationship between standards and a test appeared in the analyses:

- Test items addressing only part of the set of standards (i.e., Standards 1 and 2);
- Test items measuring only a small portion of the content objectives under the set standard; and

- The majority of the test items corresponding to only one or two objectives (mainly R2<sup>30</sup> and R32<sup>31</sup>) under a standard and only one or two items corresponding to other objectives.

In a nutshell, the Phase I results presented in Section 6.4 indicate that the rSECEE does not meet Webb's (1997) comprehensive criteria, and thus there is limited-to-no alignment between the rSECEE and the Libyan EFL content standards. For example, the rSECEE failed to meet the alignment criterion for categorical concurrence, despite the fact that it contained 60 test items.<sup>32</sup> From the number of test items corresponding to each standard, it can be argued that Libyan EFL standards were not given equal weight in the rSECEE. Although the Libyan Ministry of Education did not differentiate between standards or place greater emphasis on one standard over another, the rSECEE gave different weightings to the standards by varying the number of items measuring content related to certain standards. As seen in Table 6.6, there were 20 test items devoted to measuring Standard 1, but 38 test items devoted to measuring Standard 2. The analysis of categorical concurrence revealed that with the high number of test items being used in the rSECEE, the test developers distributed these items unevenly so that two of the six standards had 58 items measuring knowledge related to them. Furthermore, as reported by the review panel members the Libyan EFL content standards have not been adequately operationalised within the 2014/2015 rSECEE.

In addition, there was a generally acceptable level of DOK for both Standards 1 and 2. However, approximately 38% of the rSECEE test items measuring Standards 1 and 2 were at a level of complexity that was below that of the corresponding objectives. The Libyan Ministry of

---

<sup>30</sup> The student to read for detail.

<sup>31</sup> The student to develop the skill of explaining meaning in different words.

<sup>32</sup> Only 58 items were considered for analysis, because two of the items were measuring Grade 11 content.

Education set a high proportion of their EFL content objectives at DOK levels of skill concept (Level 2) and strategic thinking (Level 3). However, the test items using either MCQs, true or false format, or match were all judged to have targeted a lower level of knowledge and did not sufficiently cover the DOK levels specified in the standards. In essence, all test items were judged at DOK Level 1 that only necessitates the students simply recall information or recognise facts (i.e. low levels of cognitive domain). Therefore, I argue that 2014/2015 rSECEE test items are only recall-type questions.

Thus, the rSECEE's DOK level percentages mirror what is regularly reported in the educational literature that large-scale high-stakes tests tend to not capture meaningful aspects of students' thinking and learning and assess discrete skills instead of higher-order thinking skills such as analysing, synthesising, critical thinking, or generating hypotheses (Lane, 2004; Paris, 2000). The findings reflect Resnick and Resnick's (1989) accurately worded criticism that high-stakes tests "fare badly when judged against the criterion of assessing and promoting a thinking curriculum. They embody a definition of knowledge and skills as a collection of bits of information, and they demand fast, non-reflective replies" (p.73).

Furthermore, the Libyan EFL academic standards and the rSECEE 2015 version attained the lowest degree of alignment on ROK correspondence criterion. This may be due to the fact that the rSECEE did not cover the full range of objectives under each of the set standards. It is evident that the rSECEE did not measure the necessary breadth across the objectives with no standard meeting even the minimum threshold coverage of 50%. Given the adequate number of test items on the rSECEE, it was anticipated that items would be distributed among the objectives so that at least 50% of the objectives would have a minimum of one test item measuring content in relation to that objective. However, this was not the case. Across the

objectives of a standard, the majority of the items (58) were clustered among a few objectives rather than spanning the range of objectives. Consequently, the rSECEE can be judged to measure students' knowledge of only a small proportion of the full domain of the content knowledge stipulated by the Libyan EFL academic standards. It should be emphasized that a large-scale testing such as the rSECEE is not expected to measure the whole scope of the content objectives and curriculum. However, in order for tests to be useful, they ought to cover a wide range of standards throughout the curriculum. It is known that many teachers teach to the test, but if tests cover the full domain of the curriculum, then no harm will be done when teachers *teach to the test* (Shepard, 1991).

Possible reasons why the full range of Libyan EFL objectives went unassessed and the rSECEE attained a low alignment level on ROK correspondence could be attributed to the development process of test items. The development of test items for some objectives is more difficult and demanding than for other objectives. The Libyan Ministry of Education depends on experienced Libyan language inspectors to write the test items. The construction of effective test items is a challenging but crucial step in test development. This step is considered important because the test items are the major building block for all tests, and the methods used to create effective test items are a major source of validity evidence for any testing programme (Haladyna, Downing, & Roddriguez, 2002). It can be further argued that it is fundamental to have and employ test specifications within in a high-stakes testing context such as Libya, as it will help achieve consistent objective testing and control the production of large amounts of similar items. Therefore, test specifications within the Libyan context may become critical for: documenting the characteristics of a test to guide test construction, enhancing test objectivity, and for generating a record of evidence drawn together to address the issue of validity (Guba & Lincoln,

1989). Furthermore, as noted by Davidson and Lynch (2002), and Guba and Lincoln (1989) test specifications in contexts such as Libya may result in a record, and if this record of specification is maintained as it evolves, it can become a validity narrative. This record, as noted by Li (2006), can later be presented for peer review and as a *permanent audit trail* that can be reviewed by multiple stakeholders in Libya including: Libyan ministry officials, curriculum developers, and language inspectors.

With regards to the balance of representation criterion reported in Table 6.10, the six EFL standards were not met at the acceptable level in the rSECEE. However, Standards 1 and 2 almost met the acceptable cut-off with Standard 1 having an index of 0.650, and Standard 2 having an index of 0.521. Therefore, this suggests that rSECEE test items corresponding to Standards 1 and 2 are weighted heavily towards certain objectives and are not evenly distributed across the hit objectives. When the index is near zero, the test is not balanced. An unbalanced assessment, as noted by Bhola et al. (2003, p.24), does not provide “scores that readily facilitate unbiased inferences being drawn”, because the construct may have been underrepresented, or one standard was over weighted, as was the case with the rSECEE. Therefore, from the balance of representation index, the documentary analysis data, and the responses from the panel review members and test developer, it can be argued that the rSECEE does not accurately measure the target construct. By assessing a limited portion of the target construct the rSECEE is grossly underrepresenting the construct, thus there is a case of CUR (Messick, 1989).

In addition, the rSECEE having test item flaws, (such as include ambiguously worded test items, questions without a stem, tricky items, and spelling errors), may incorrectly inflate test scores of Grade 12 Libyan students who actually may not know the information tested by the question, and thus contribute to CIV. Furthermore, the rSECEE comprising of items testing too

much information and not reflecting a specific content offers unintentional cues that could facilitate students eliminating incorrect options. Consequently, it may be testwiseness<sup>33</sup> rather than learning that determines whether Libyan 12 Grade students choose the right answer. In addition, tricky items where the discrimination among options were too fine and have window dressing (irrelevant extraneous material) may, as cautioned by Dodd and Leal (2002), deceive test-takers into choosing the distractor instead of the right answer and may even increase students' test anxiety. A partial sentence item format is also problematic because test-takers have to hold the partial sentence in their working memory and consecutively complete it with the correct answer (Statman 1988). Therefore, it can be argued that inaccurate and confusing test items may erroneously lower Libyan students' SECEE test scores.

Students correctly answering MCQs by guessing can also augment CIV for the rSECEE. The MCQs had a fixed set of four answer options; the so-called Goldilocks formula (Hubley, 2012). Therefore, there is a probability of getting the correct answer by guessing (Downing, 2002). With the Goldilocks formula the test-taker has a 25% chance of randomly guessing the correct answer. It can be further added that argued that if a Grade 12 Libyan student's ability level is below the set cut-off passing score, "successful guessing on poorly crafted MCQs may provide a sufficient number of positive score points to erroneously pass him/her" (Downing, 2002, p.237).

Moreover, with reference to all the rSECEE test items being at DOK Level 1 it could violate the AERA (2000) position statement for high-stakes testing in pre-K-to-Grade12 education. The position statement emphasizes that "[b]oth the content of the test and the cognitive processes engaged in taking the test should adequately represent the curriculum. High-stakes tests should not be limited to that portion of the relevant curriculum that is easiest to measure" as with the

---

<sup>33</sup> Testwiseness refers to a set of behaviors that allows examinees to maximize their test score (Downing, 2002, p.237).

rSECEE. The 2014/2015 rSECEE has limited its test items to the recall of vocabulary and discrete and isolated grammar components because “tests of grammar rules and of translations are easy to construct and can be objectively scored” (Brown, 2000, p.19). Although the rSECEE multiple-choice responses may provide some information about Libyan students’ Grade 12 content knowledge, as a one-time paper and pencil assessment it has “serious limitations in measuring the variety and scope of classroom learning” (Marchant, 2004, p.3).

With the rSECEE consisting of only discrete-point testing items it can be further argued that the test is mirroring a structuralist approach to language learning and testing. Structuralist discrete point language tests based on Lado’s positivistic perspective were common in the 1960s (Farhady, 2018). According to his model, “language consisted of sounds, words, and sentences manifested in the four language skills” (Farhady, 2018, p.2). In addition, language ability is perceived to be the “sum of the knowledge of an individual on these components and skills” (Farhady, 2018, p.2) and thus language tests under Lado’s framework ought to be “measuring these components component, assuming that the sum of the performance on these components would be an indication of the learners’ overall language ability” (Farhady, 2018, p.3).

Structuralist discrete point language tests are designed to measure language ability with each item tapping into one particular element of language and at the same time covering as many elements of language components and skills as possible (Farhady, 2018). The common test item format under this model is the multiple-choice format, because of its “ease of scoring and making the testing process practical” (Farhady, 2018, p.3). However, the structuralist discrete point test design does not correspond to the realistic use of language in authentic contexts, or to the Grade 12 EFL content objectives which aim at promoting Libyan EFL students’ communicative competence. In addition, the rSECEE testing practices is inconsistent with

research about the models of learning and is decontextualized. The rSECEE was considered to be decontextualized because, as Resnick and Resnick (1989) rationalised, it lacks authentic purpose presuming that skills are stable regardless of time and reason and breaks language into small discrete point elements. The rSECEE breaks the measured language skills (grammar, reading, and vocabulary) into small isolated discrete elements rather than measuring all language skills in an integrated and cohesive manner.

From the document analysis and the Webb (1997) alignment criteria results it can be further argued that the rSECEE test items measuring reading skills only consider language as a linguistic competency that it intends to measure, rather than language being a means to an end; something to be used in order to understand a reading text (see Table 6.10 for examples of rSECEE reading test items). Such an understanding ought to include understanding the main idea of the text, interpreting the text, relating the text to its context and the outside world, making judgments about the text, evaluating and synthesising the text, and so on. In other words, language is a means to an end of communicating at high levels of cognition.

Table 6.10

*Examples of the rSECEE Test Items Measuring Grade 12 Students' Reading Skills*

| Test item Number | Item Stem   | Target Objective Being Measured                         |
|------------------|---|---|
| 1.               | Bronze was first made in the middle east.<br>A. True<br>B. False                        | R2: To read for detail.                                 |
| 4.               | Early learning should continue to be rewarded from time to time.<br>A. True<br>B. False | R17: To practice reading for global meaning.            |
| 5.               | A person who watches a sporting team is a spectator.<br>A. True<br>B. False             | R10: To work out the meaning of phrases from context.   |
| 10.              | American people use the present perfect tense less often.<br>A. True<br>B. False        | R20: To raise awareness of different styles of writing. |
| 44.              | Our bodies tell us ..... we need.<br>A. what<br>B. who<br>C. whom<br>D. where           | R31: To raise awareness of reference words.             |

The rSECEE test items devoted to reading measure examinees' linguistic abilities without reference to higher order thinking skills. The reading test items were measuring thinking that only emphasizes recall, memorization, and identification, which is at a lower level of thinking (i.e. DOK Level 1). From a cognitive perspective, reading comprehension can be described as the ability to create linguistic meaning from written representations of language. This ability depends to a great extent on two equally important competencies: language comprehension and decoding (Hoover & Gough, 2013). The former is the ability to construct meaning from spoken representations of language, and the latter is the ability to recognise written representations of words (Hoover & Gough, 2013). Both language comprehension and decoding are essential for reading comprehension success. Neither competency is adequate in itself. For example, being

fully competent in a language but having no ability to identify its written words will not allow for successful reading comprehension, neither does having the ability to recognise the written words of a language but not having the ability to understand their meaning. Weakness in either ability results in weak reading comprehension (Hoover & Gough, 2013). Despite the essential importance of assessing both language comprehension and decoding abilities when measuring language learners' reading competency, they are not incorporated within the rSECEE test items devoted to the reading skill. The absence of a reading text and the reading test items being small isolated discrete point elements means that the rSECEE is not adequately measuring Libyan Grade 12 students' attainment across all six standards. There seems to be "no interest in the thought processes, only in the performance" (Baird et al., 2017, p.319).

Tests covering particular content can be at the expense of other content which then goes untested and thus will not be taught (Dillon, 2006; Jerald, 2006; Tienken, & Wilson, 2001). For instance, the rSECEE covering grammar and reading in isolated discrete point items is at the expense of other skills such as speaking, listening and writing, and are possibly not taught. In determining the scope of assessment, it should be recognised that "[t]esting cannot be neutral on what is taught and learned. Any test is an expression of values on teaching and learning" (Cole, 1999, p.1). In this context, memorization and basic recall of information holds prominence within the Libyan culture of learning which perceives education as a process of conveying information. With the 2014/2015 rSECEE echoing the importance of memorization, it is supporting Al-Buseifl (2003) and Alhmali's (2007) criticism of the high-stakes examination system, particularly the rSECEE for its tendency to test rote memorisation (see Section 2.5). In addition, the rSECEE's emphasis on memorization is not ensuring that meaningful aspects of students' thinking and learning are being captured. Given that assessments are intended to

evaluate and measure the learning that was supposed to have taken place in the classroom, a more accurate and rigorous conceptualisation of the Grade 12 EFL construct may produce a better high stakes test and, thus, have higher validity (Baird et al., 2017).

Considering the high stakes of the rSECEE within the Libyan context, there is a fundamental need to support the relevance of the inferences made. If the Libyan Ministry of Education is to test students, there is a need to “collect and present hard evidence that the test measures what is intended and that the inferences drawn from test scores are more-or-less accurate and more-or-less defensible” (Downing, 2002, p.239). As noted by Downing (2002), Downing and Haladyna (2004), and Haladyna and Downing (2011), the issues that bring about CIV and CUR are undoubtedly under the control of test developers; in the context of this dissertation, they are under the control of the selected Libyan EFL language inspectors. The rSECEE test developers need to put greater effort into constructing test items that have higher DOK level items, items measuring useful content knowledge representing the Grade 12 EFL content objectives, and ought to avoid ambiguous and tricky items that can lead to both guessing and cueing, and hence reducing CIV. If test scores do not accurately support the valid inferences about the subject matter being assessed, it may also result in test-score inflation. Test-score inflation can also be brought about by coaching and cheating on the part of the teacher, including the teacher directly assisting students during the administration of high stakes tests, or in extreme cases answering questions on their behalf (Jacob & Levitt, 2003). Another factor that contributes to test score inflation is the use of the same test format year after year. As argued by Baird et al. (2017, p.323), there may be a substantial overlap of 40% or more of test items from one year to another. Thus, Libyan student achievement increases may be due to “differences in familiarity with the

specific test content, and therefore may not generalize to the broader domain of achievement” the rSECEE is meant to represent.

In a nutshell, it can be argued that there is limited to-no- degree of alignment between the rSECEE and Libyan EFL content standards; hence a weak link exists within the chain (Näsström, 2008). The components within the Libyan education system may have drifted apart and may even possibly be giving totally different messages to both teachers and students about what they are expected to know and be able to do. As a result of the weak alignment, the Grade 12 Libyan education system is operating less effectively, and the basis of any decisions taken will also be weak. There is also the chance that its components are emphasizing different knowledge and skills, contradicting each other, isolating from one another and may eventually break, and hence, potentially increase the level of anxiety and pressure for both teachers and students.

Furthermore, in an unaligned system such as the Libyan secondary level education, students, parents, educators, administrators and policymakers may be misled by reports and scores inferences. This is because the rSECEE results may not signify that students have attained the articulated standards and consequent findings may not respond to deliberate actions by students and educators (Baker, 2005; Herman, 2004; Webb, Herman, & Webb, 2007).

Without a semblance of alignment, nothing will cohere in an educational system (Biggs & Tang, 2001; Biggs, 2003; Näsström; 2008) and with negative consequences for the whole education system including negative washback. When testing is for school accountability or to influence the curriculum, the test should be aligned with the curriculum as set forth in the standards’ documents setting out the intended goals of instruction, because high-stakes testing inevitably creates incentives for inappropriate methods of test preparation (West, 2010). As a

very small sample of the EFL standards were assessed in the rSECEE there may be a risk, which has been strongly emphasized at a general level by Resnick et al (2004), that Libyan EFL teachers may be only focussing their EFL classroom language instruction on the standards that are assessed and totally ignoring the unassessed standards. Thus, the link between standards and teaching is weakened, and the students may not be getting the optimal chance to attain all the standards, and hence a possible negative washback. In other words, there may be a possibility that the rSECEE may “influence [Libyan] language teachers and learners to do things they would not otherwise necessarily do” (Messick, 1996, p. 241).

Moreover, Heubert and Hauser (1998) argue that the significance of a test does not lie in its validity in general, “but its validity when used for a specific purpose” (p.3). Accordingly, in line with Heubert and Hauser (1998), if the rSECEE is valid for influencing classroom practice, driving the curriculum, or holding schools accountable it may not be “appropriate for making high-stakes decisions about individual student mastery” unless the curriculum, the teaching, and the test are aligned (p.3). Based on the reported findings and analysis, the Libyan Ministry of Education and, in particular, the test developers need to reconsider the DOK levels for the unmeasured standards and issues of assessing complex thinking and the narrowing of the curriculum in their rSECEE test design before conclusions are made about the Libyan Grade 12 students’ attainment of the standards. Considering that there is limited to-no- alignment between the Libyan EFL content standards and the rSECEE and that the lack of alignment between components of an education system may bring about negative washback (Linn, 2000; Shohamy, 1997; Tan, 2008; Wall, 2005), it could be presupposed that the Grade 12 EFL classroom is characterised by: teaching activities that indicate negative washback; increased attention on grammar and vocabulary learning; little or no focus on the listening, writing, or speaking skills; a

reduction in the amount of instruction and a narrowing of the curriculum; and learners who pay a great deal of attention to grammar and vocabulary learning. These assumptions can only be confirmed through empirical evidence, and, thus, Phase II investigates the nature and scope (Cheng, 2004) of the washback (if any) of the rSECEE on the Libyan EFL Grade 12 classroom. In addition, the empirical evidence derived from Phase II is to help identify the type of relationship between the degree of alignment and the washback of the rSECEE. Thus, Chapter Seven presents a detailed description of the data collection and analysis procedures for Phase II along with analysis, strategies, results, and a discussion of the results.

## Chapter VII

### Phase II Study

*...I don't want a textbook, I want language and how to use language. That's all I want. So, fix the test...* (Student comment, Focus-Group 2)

As discussed in Chapter Four, high-stakes tests in educational systems around the world have had many shortcomings (Madaus, 1985), such as negative washback on teaching and learning (Alderson and Hamp-Lyons's (1996, Shohamy et al., 1996). The shortcomings cited in the educational and language assessment literature include examples of high-stakes tests which: (1) fail to assess students on what they were taught in the classroom (Fox & Cheng, 2007); (2) cause a narrowing of the curriculum (e.g. Amrein & Berliner, 2002; Cimbricz, 2002; Shepard, 2002; Smith, 1991a); (3) affect students' health and well-being (e.g., Polesel et al., 2012); and (4) constrain teachers' abilities to meet the sociocultural needs of their students (Roach, Niebling & Kurz, 2008). It was also noted in Chapter Three that researchers have suggested that the lack of alignment between standards, curriculum and testing may lead to negative washback (Green, 2007, 2014; Linn, 2000; Shohamy, 1997; Tan, 2008; Wall, 2005).

The findings from Phase I of the study indicated that there is insufficient alignment between the rSECEE (an under-researched, life-changing, high stakes test) and the Libyan EFL standards. This encouraged me to investigate the nature and scope of the washback (if any) of the rSECEE on teaching and learning in Libyan EFL Grade 12 classrooms. Phase II of this dissertation research explored washback on teaching and learning in EFL classrooms within this context of *misalignment*. In so doing, I began to address the proposition of those researchers who have argued that misalignment leads to negative washback. Subsequently, the findings of Phases I and II are synthesized to arrive at an interpretation of the relationship between the degree of

alignment and the washback of the rSECEE and the possible implications of this relationship for key stakeholders, such as policy-makers, test developers, teachers, and students.

### **7.1 Research Questions**

The findings from Phase I indicated that rSECEE does not meet Webb's (1997) comprehensive criteria for alignment. Indeed, the findings lead convincingly to the conclusion that the rSECEE is misaligned with the Libyan EFL content standards. This formed the backdrop for the third research question, namely:

**What is the nature and scope of the washback (if any) of the rSECEE on Libyan EFL Grade 12 teaching and learning?**

This overarching question was addressed by exploring the following sub-questions:

- g) What evidence is there of washback of the rSECEE on EFL teachers, i.e. how does the rSECEE influence teachers' accounts of teaching and testing (i.e., both external and internal testing)?
- h) To what extent does the rSECEE appear to influence these teachers' teaching practices?
- i) How does the rSECEE influence EFL learners' accounts of learning?

In accordance with Cheng (2004), teachers' accounts are operationally defined in this study as teachers' comprehension and understanding of aspects of classroom teaching in relation to the rSECEE. Aspects of teaching that were the focus of investigation included teachers' accounts of:

- 6. The reasons behind the change in examination format and content;
- 7. The test format;
- 8. Any necessary extra work or pressure in teaching towards rSECEE;
- 9. Any changes in teaching methodology employed; and

#### 10. Any challenges whilst teaching.

In addition, students' accounts were operationally defined as students' knowledge and understanding of the underlying principles of the rSECEE, and students' views about the rSECEE. Learning practices were operationally defined as what students say they do in order to prepare themselves for the rSECEE and to improve their level of English.

## **7.2 Method**

### **7.2.1 Research Setting**

Based on Cheng and Curtis's (2004) and Watanabe's (2004) emphasis on the importance of providing a detailed description of the research context, Chapter Two provided a description of the macro context and the role the rSECEE plays within the Libyan context. The following section provides a description of the micro context of Phase II, which took place within two high schools. The research sites and the participants have been given pseudonyms to protect their identities. I have used pseudonyms UM and AL to represent the two schools.

#### **7.2.1.1 Description of High Schools UM and AL**

The two high schools where this study took place are located in Libya's third largest city. UM high school consists of 15 classrooms with a total of 273 students, of whom 70 were enrolled in Grade 12 in the 2016/2017 academic year, during which the Phase II data were collected. UM high school is located in the centre of the city but teachers and students from other neighbourhoods can attend it, because in the present research context, students can attend schools of their choice. In contrast, AL high school is a female only school located in a suburb consisting of 17 classrooms with total number of 475 enrolled students during the 2016/2017 academic year, of whom 132 were Grade 12 students. The majority of the students attending and teachers teaching at AL high school were residents of the suburb or nearby suburbs. Similar to

other high-schools within Libya, both UM and AL high schools offer the General Secondary Programme for the literary and scientific sections. The total number of teachers teaching at both schools during the academic year 2016/2017 is summarised in Table 7.1, while the rSECEE's students pass rates over the past three years at AL and UM high schools are displayed in Table 7.2 and Table 7.3, respectively.

Table 7.1

*Numbers of Teachers in UM and AL High Schools*

| <b>School</b>    | <b>Total Number of Teachers</b> | <b>Total Number of Grade 12 Teachers</b> | <b>Total Number of Grade 12 EFL Teachers</b> | <b>Number of Teacher Participants</b> |
|------------------|---------------------------------|--|--|---------------------------------------|
| <b>UM School</b> | 112                             | 17                                       | 2  | 1                                     |
| <b>AL School</b> | 136                             | 20                                       | 3  | 2                                     |

Table 7.2

*AL High School Students Pass Rates in rSECEE from 2015-2017*

| <b>Year</b> | <b>Total Number of Students</b> | <b>Total Number of Students Passing the rSECEE 1<sup>st</sup> Sit</b> | <b>Total Number of Students Failing the rSECEE 1<sup>st</sup> Sit</b> | <b>Total Number of Students Passing the rSECEE 2<sup>nd</sup> Sit</b> |
|-------------|---------------------------------|---|---|---|
| <b>2015</b> | 109                             | 109   | 0   | 0   |
| <b>2016</b> | 93                              | 87  | 6   | 6   |
| <b>2017</b> | 132                             | 121   | 11  | 11  |

Table 7.3

*UM High School Students Pass Rates in rSECEE from 2015-2017*

| <b>Year</b> | <b>Total Number of Students</b> | <b>Total Number of Students Passing the rSECEE 1<sup>st</sup> Sit</b> | <b>Total Number of Students Failing the rSECEE 1<sup>st</sup> Sit</b> | <b>Total Number of Students Passing the rSECEE 2<sup>nd</sup> Sit</b> |
|-------------|---------------------------------|---|---|---|
| <b>2015</b> | 117                             | 108   | 9   | 9   |
| <b>2016</b> | 95                              | 84  | 11  | 2   |
| <b>2017</b> | 53                              | 39  | 14  | 5   |

Although it is very difficult to argue for the ‘representativeness’ of any school in Libya, both UM and AL high schools can be considered to be typical secondary schools in terms of the infrastructure and the numbers of teachers and students at the schools. The context for Phase II was selected for both theoretical and personal reasons.

The theoretical rationale underpinning my choice of these research sites included the lack of research on the washback of rSECEE, despite the fact that it is a high stakes examination that determines if students graduate from high school and to which university or college they will be admitted. Therefore, the Grade 12 high school classroom offers a suitable research context in which to investigate washback from rSECEE at the classroom level. In addition, a similar context was researched by Onaiba (2014), who conducted a study about the washback effect of a revised BECEE on teachers’ instructional practices, materials and curriculum. By combining Onaiba’s findings (2014) and the findings of the current study, a clearer picture can be obtained about the washback of the revised test in Libya. Moreover, the evidence should also enrich the wider understanding of the washback phenomenon.

The research context was also chosen for personal and practical reasons. I used to teach at the UM high school and the recruited classroom observer had taught at the school for more than ten years, which was a key benefit for observing the Grade 12 EFL classrooms and eliciting richly informed observations from the interviews.

## **7.2.2 Participants**

### **7.2.2.1 Teachers and Students**

As highlighted in Chapter Five (Section 5.6), Phase II was a qualitative study that elicited data through the use of a questionnaires, classroom observations, and semi-structured interviews with three Libyan Grade 12 teachers. The main participants in this Phase were teachers and

students. However, the two policy-makers' interviews that were part of Phase I were also considered in the analysis and interpretation stage, as these helped to clarify and validate the findings of Phase II.

Three Libyan EFL teacher participants ( $n = 3$ ) were identified by applying a purposive/judgemental sampling technique (Tashakkori & Teddlie, 2003) with three parameters: they had taught the Grade 12 English curriculum for at least one year and whilst the rSECEE has been in operation; they had experience in teaching other secondary grades (Grade 10/11); and they were from different geographical locations within the research context. In addition, variations in participants' ages, teaching experience, and degree of education were also taken into consideration. Table 7.4 summarises the three teachers' background information and classroom characteristics collected from their questionnaire responses and the observation of their EFL classrooms.

Table 7.4

*Teacher Demographics and Classroom Characteristics*

| <b>School</b>   | <b>AL</b>  |          |            |          | <b>UM</b>  |          |
|---|------------|----------|------------|----------|------------|----------|
| <b>Teacher</b>  | <b>A</b>   |          | <b>B</b>   |          | <b>C</b>   |          |
| <b>Age Range</b>  | 35-45      |          | 25-35      |          | 25-35      |          |
| <b>Current Degree</b>                                     | B.A.       |          | B.A.       |          | B.A.       |          |
| <b>EFL Teaching experience</b>                            | 5-10 yrs.  |          | 5-10 yrs.  |          | 5-10 yrs.  |          |
| <b>Grade 12 Teaching Experience</b>                       | 12 yrs.    |          | 7 yrs.     |          | 7 yrs.     |          |
| <b>Date of Observation</b>                                | April/May  |          | April/May  |          | April      |          |
| <b>Section</b>  | Scientific | Literary | Scientific | Literary | Scientific | Literary |
| <b>Class Size</b>   | 33         | 27       | 25         | 0        | 37         | 15       |
| <b>Number of Students who Failed the rSECEE First Sit</b> | 1          | 0        | 1          | 0        | 2          | 1        |
| <b>Seating Arrangements</b>                               | IR         | IR       | IR         | IR       | IR         | IR       |

Note: B.A. = Bachelor of Arts degree, IR = In rows. Number of students who failed the first sitting of their rSECEE was for the academic year 2016/2017.

Another participant in this Phase was Teacher HM, who was a former teacher at the UM high school, and is currently a graduate student. Teacher HM had more than ten years of Grade 12 EFL teaching experience. Because I was away from the research site during part of the data collection period I asked Teacher HM to conduct the observations on my behalf. Prior to the classroom observations, Teacher HM was trained to use the observation tools. The observer was also informed about the research purpose and the Phase I findings. In addition, I gave HM a brief overview of the washback phenomena and the importance of observations in gathering the necessary data. Details of steps taken to ensure the reliability of HM observational coding are provided below (see Section 7.6.2).

### **7.3 Data Collection**

Having briefly described the data collection instruments and procedures in Chapter Six, this section provides further detail on each tool and the four stages of data collection (see Figure 7.1 below).

#### **7.3.1 Instruments**

I employed two questionnaires, observations, semi-structured interviews, and focus group interviews with students to help explore how washback from the rSECEE plays out in the Grade 12 Libyan EFL classroom. The combination of instruments helped me understand the complex nature of washback.

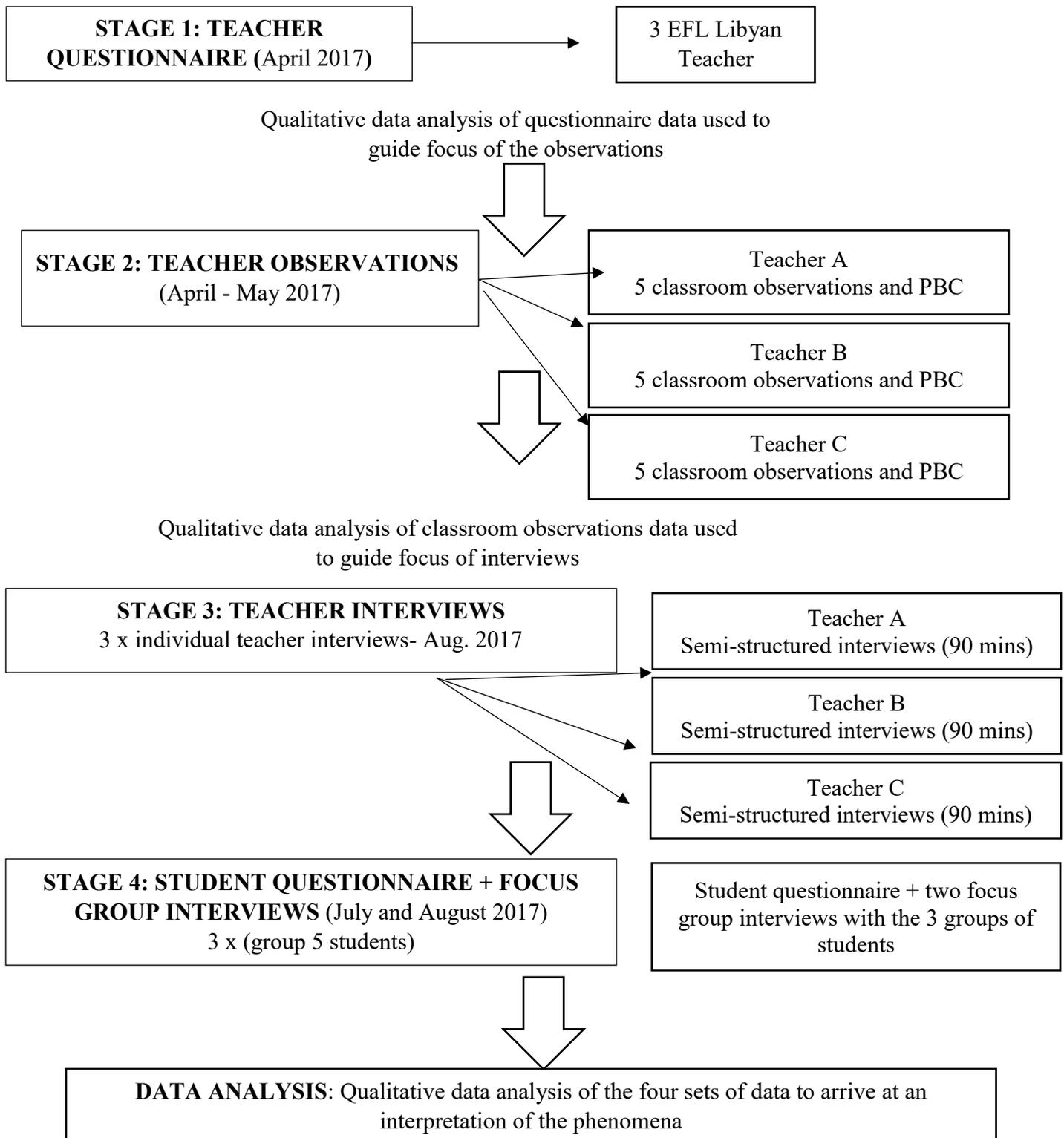


Figure 7.1: Phase II Data Collection and Analysis

Source: Adapted from (Booth, 2012)

### **7.3.1.1 Stage 1: Questionnaires**

Before distributing the teacher questionnaire, I contacted six teachers by both phone and e-mail during the month of March 2017 to invite them to participate in my research. However, only three responded positively during the month of April (see figure 7.1). Distributing the questionnaire at the start of the study proved useful, as it provided the participants with a framework for reflecting upon concepts that were the focal point of subsequent stages. Not only was it a spring board for thinking for the participants, but it guided me in leading the conversation to elicit answers to my main research question. More importantly, the patterns identified in the questionnaires data provided a stimulus for teachers semi-structured interviews.

### **7.3.1.2 Stage 2: Observation**

Classroom observations were employed because it was believed that they could provide me a “better understanding of the participants’ actions and the reasons behind them” than could be obtained from solely questionnaires or interviews (Wall & Horák, 2007, p. 106). Indeed, through observation and observation recordings I was able to gain a deeper understanding of how both teachers and students made sense of rSECEE from their classroom behaviours. The observations took place from April to early May 2017. The washback literature documents the relationship between the time of the academic year and the intensity of the washback (Cheng, 1997; Shahamy et al., 1996; Watanabe, 1996). Therefore, the observations were purposely undertaken during this period because the SECE takes place during the month of July, and intensive exam preparation normally starts end of May or early June in Libya. Thus, the observed classes should represent classroom norms in terms of Grade 12 EFL language teaching and learning practices. According to Cheng (1997), normal conditions for both teaching and learning may not be evident at either the very beginning, or at the end of school year.

From a methodological point of view, observing every class over an academic term or year would have been valuable, as it would have allowed me to obtain a more accurate picture of the Grade 12 EFL classroom and recorded the possible variations. However, for practical purposes a sampling approach was employed for the observations.

As explained earlier, I was unable to observe the EFL classes in person, therefore, I advised Teacher HM to keep her notes detailed in order to capture as much information as possible about the classroom procedures that I was unable to observe. I also asked her to pay specific attention to the classroom activities, any feedback provided by the three teachers to their students, the type of teacher-student interaction, and any classroom episodes that could be linked to or encouraged by the revised Grade 12 EFL curriculum.

The observation tools and procedure were piloted over two Grade 12 EFL classes. The purpose of piloting the observation tools was two-fold. The first reason was to validate observational tools such as the note taking observational matrix. The second reason was to help check the clarity of the instructions given to Teacher HM and the rigorousness of her observational notes. The data collected from the two pilots was peer-reviewed by my research supervisor and her suggestions were incorporated when revisions to the observational tools were made.

Teacher HM observed five classes for each teacher, making a total of 15 classes comprising 675 minutes of Grade 12 instructional time. To record the observations, I employed Watanabe's (2004) note-taking coding matrixes (see Appendix P). In accordance with Watanabe (2004), various classroom events were recorded on a note-taking coding matrix. This coding matrix consisted of very broad categories such as "time", "material", "what teacher is saying/doing", "what students are saying/doing", "what is written on chalkboard", and "the observer's

comments and questions”. With the teachers’ permissions all the observed classes were audio-recorded.

After each classroom observation HM would send me the audio recording along with the notes and we would discuss the recording over Skype and correct any gaps that I may have misinterpreted. After the first couple of observations, I felt Teacher HM became very familiar with the observation process and tools, I therefore asked her to focus more on how the teachers were using the textbooks. In particular, I asked her to concentrate on if the teachers were: using the books according to the Grade 12 EFL content objectives of which Teacher HM has sound knowledge; operationalising these content objectives in their daily classroom practices; giving each language skill equal weighting as was emphasized in the Grade 12 EFL textbooks; or just working on the textbook content that only resembled the rSECEE content. If the teachers were not operationalising the Grade 12 content objectives in their pedagogical practices, it was important to understand the rationale behind the teachers’ choices of classroom content, and the rationale for favouring certain language skills over others. Therefore, these were a spring board for the post-observation chats and semi-structured interview questions.

After each observation, I asked the three observed teachers about their objectives with regard to certain activities and episodes that took place during the lesson. These post-observation chats took place in Arabic and were later translated into English. Importantly, to gain a more accurate picture of the Grade 12 EFL Libyan classroom, it was also necessary to document on COLT A (a description of this tool is provided in appendix Q) further details on any rSECEE information provided by the teacher; these were analysed separately. In accordance with Alderson and Hamp-Lyons (1996, p. 288), these activities included time available for students to talk as a percentage of class time, time spent on pair-work as a percentage of class time, number of references to the

rSECEE per lesson, frequency of metalanguage use per lesson, average number of ‘innovations’ per lesson, and average frequency of shared laughter per lesson. The start time for each activity was recorded and the duration of each episode and as a percentage of the total class time were also calculated. The lessons were coded, and the results were calculated and combined for each of the observed lessons.

### **7.3.1.3 Stage 3: Semi-Structured Interviews**

After processing the results of the teacher questionnaires and observations, I conducted the three interviews with the three Libyan teachers. The purpose of these interviews was to extend and validate initial findings by encouraging each participant to talk extensively about their classroom experiences, while making sure that the interview did not ‘turn off track’. I gave each participant ample flexibility to talk about whatever she wished to articulate. It is often the situation that one teacher has constructive insights into one particular element of being a language teacher and classroom practices, while a different teacher has similarly important insights into a different element (Senior, 2006). As many teachers would do, the three participants contributed insights gained from their current Grade 12 teaching practices in Libya, as well as providing insights from their language teaching experiences in other contexts, their personal experiences as language learners and graduate students, and their personal life experiences.

The insights offered by one teacher produced further questions that the next teacher could be asked. This form of questioning is known as ‘theoretical sampling’ within the grounded theory approach (Charmaz, 2006). This particular technique enabled me to explore the implications of additional phenomena that may not have been regarded as important before undertaking the research. Furthermore, by “asking [a] subsequent teacher[s] to elaborate on insights and

observations provided by previous teachers, [allowed for the development of] a compromise picture of what all the teachers are collectively saying” (Senior, 2006, p. 21). Besides the “validity of the findings is also enhanced, since the researcher can check whether the insights provided by one teacher are unique or shared by others” (Senior, 2006, p. 21).

For each interview, I put the teachers at ease by asking them about their past teaching experience, and then asking them for their personal insights on the revised testing system and in particular the rSECEE. In terms of technique, I ensured that I asked carefully thought out open questions rather than leading ones in order to “avoid blurting out loaded questions and to avert forcing responses into narrow categories” (Charmaz, 2006, p. 18). My questions asked the participants to describe and reflect upon their experience in particular ways that rarely take place in everyday life. I also frequently encouraged the participants to expand on each point they made using a questioning voice. This meant that they further explained a particular incident or presented an example of what they meant by a particular concept. I also used ‘devil’s advocate’ style of questioning, along with hypothetical questions, such as ‘what if?’. For example, Teacher C said that she implemented communicative approaches in her classroom techniques and was not teaching to the test. In order to understand what she actually meant by communicative approaches and teaching to the test, I asked her to give an example of what she does in her classroom that she believed was communicative.

In addition, I encouraged the participants to describe events and behaviours that they believed involved necessary extra work or pressure in teaching towards the rSECEE, any changes in teaching methodology they employed, or any challenges whilst teaching to the new curriculum. After their clarifications and responses, I provided on-the-spot validity checks, by repeating what they had said during the interview and also to ensure that I had understood them correctly.

As each interview progressed and themes in the data started to emerge, I considered secondary sources such as washback and high-stakes testing research literature (see Chapter Four). Insights from these sources further informed my understanding of the teachers' accounts of the rSECEE and its washback on teaching and learning. Teachers A and B's interviews were conducted only in Arabic; however, with Teacher C, I was code-switching between English and Arabic as she spontaneously replied in English. All three interviews were audio-recorded, and two lasted approximately 90 minutes, while Teacher C had so much to say that her interview lasted for 120 minutes. The recordings were transcribed and translated into English by myself and checked for accuracy by Teacher HM.

#### **7.3.1.4. Stage 4: Focus Groups**

I felt it was important to consider Grade 12 students' accounts of the high stakes test because, as noted by Barr (2016), students' views about the school and the classroom climate have a major impact on learning, motivation, satisfaction, and achievement. To the best of my knowledge, no researcher has examined students' accounts of the rSECEE and no feedback from test-takers and test-developers has been collected or reported.

With the Principals' approvals and before the end of the last observational class of each teacher, HM asked the EFL teacher to leave the class and invited the students (in Arabic) to participate in my research. At my request, she explained to the students the overall purpose of the research and stressed the importance of their participation and highlighted how their participation could help educators understand the impact of the rSECEE and inform Libyan policy-makers about the effects of their new curricular implementations. A detailed description and step-by-step description of the focus group process was given to the students. It was strongly emphasised that their teachers would not be informed of their responses. Letters of information were distributed

to the students and those who were interested in participating in the research had to email me attaching their consent form. To ensure randomness, I decided that the first five students from each teacher to contact me would be chosen for the focus group interviews. A total of 15 students ( $n= 15$ ) participated in three groups of five according to each of the three teachers.

Overall, the focus groups were homogenous in terms of EFL academic background but were heterogenous in terms English competency level. As noted by Crabtree and Miller (1999), homogeneity in focus group interviews can be advantageous because of the familiarity that comes with the shared background or experience, which, in turn, can help in facilitating open communication and the exchange of ideas. It was noticed in the second focus group interviews that the students were more comfortable and felt a sense of safety in expressing their concerns.

The focus group interviews took place before and after the students took the rSECEE. Therefore, there were six focus groups interviews. Before commencing with the first focus group interviews, participants were asked to complete a questionnaire which was piloted for the same reasons as the teacher questionnaire (see Chapter Five). The questionnaire was piloted by a sample of five Grade 12 students. If students had a problem understanding the questions they were orally translated in Arabic. On average the student questionnaire took less than ten minutes to complete. More importantly, the patterns identified in the questionnaires data provided a stimulus for the first set of student focus group interviews.

During the focus group interviews, participants were prompted with a series of questions to initiate a group discussion (see Appendix P). The first focus group interview with the three groups was only one hour long, while the second series lasted for one and one-half hours. The students were not asked to volunteer sensitive personal information and were not exposed to unpleasant or contentious questions. They were informed upfront that if they felt uncomfortable

about answering any of the questions, they had the right not to answer. All six focus group interviews were conducted in Arabic and with the students' permission were audio recorded. The recordings were transcribed and translated into English by myself and checked for accuracy by HM.

Importantly, to further clarify any unclear points, or to confirm or refute certain findings, follow-up contact was made with teachers and students either by phone or email during and after the data collection. In addition, to determine the confirmability and dependability of the study inquiry, *auditing* was undertaken. This involved the creation of an audit trail that consisted of raw data, data reduction, analysis products and so on. The audit was examined by the research supervisor and Teacher HM to confirm my findings and the interpretations of those findings. In each of the data collection procedures, I attempted to assess the fit between my early research curiosity and awareness and the data, without forcing my predetermined notions and hypotheses on the data.

#### **7.4 Analysis**

The Phase II study followed a qualitative data analysis approach in which qualitative data were gathered from four different sources: questionnaire; teacher classroom observations; semi-structured interviews with three Grade 12 EFL teacher; and six focus group interviews with 15 Grade 12 students. These data sets generated 11.25 hours of teachers' classroom observations, audio recordings, five hours of teacher interview audio recordings, and 11.5 hours of focus group audio recordings; a total of almost 30 hours.

Following Turner (2009), semi-structured interviews, focus group interviews, classroom observations and post-observation chats were analysed qualitatively to identify categories that were relevant to my initial washback presuppositions and research questions. The data analysis

was an ongoing and iterative process in which data collection and analysis informed each other (Mertens, 2005). This required me to delay the observations from Stage 2, until I had completed the analysis of all of the questionnaire responses; the interviews from Stage 3 were not conducted until the observations were analysed, and the second interviews were not conducted until the first interview was analysed. This procedure supports Stake's (1995) notion that "[t]here is no particular moment when data analysis begins. Analysis is a matter of giving meaning to first impression as well as to final compilations" (p. 71). When examining participants' responses from the questionnaires, interviews, or focus groups, I took Wood's (1996) key points concerning data:

The importance of the issues to the teachers is signalled by the frequency of their occurrence, their centrality with regard to other issues, and by explicit mention, by tone of voice and other signals of highly loaded issues, and other means of evaluation. The relationships among themes are signalled and can be deduced by the way in which the themes are embedded in sentences and contexts which include mention of other themes. (p. 32)

In order to minimise any systematic bias, data sets gathered from the four sources were compared or triangulated with data from a different source. As noted by Dorneyei (2007), "if we come to the same conclusion about a phenomenon using a different data collection/ analysis method or a different participant sample, the convergence offers strong validity evidence" (p. 61).

How each set of data gained from the four instruments was analysed is discussed below.

#### **7.4.1 Questionnaires**

The small size of the teacher sample ( $n = 3$ ) did not allow for any statistical analysis. The three teachers' demographics and professional experience, as well as their initial accounts on the

revised secondary level EFL curriculum and the rSECEE gathered from the questionnaire were summarised. In contrast, the student sample ( $n = 15$ ) permitted basic statistical analysis with the results quantified and presented in bar charts (see Section 7.5).

#### 7.4.2 Observations

The data sets collected through the 675 minutes of observations allowed for frequency analyses, using Watanabe's (2004) observation coding sheet A and B. These coding sheets were modelled on the COLT A observation scheme (for an example of a coding sheet used in this study see appendix S). I listened to each recoding several times in order to accurately code the observations on the coding sheets. Similar to Watanabe (2004, pp. 135-136), the following categories and their sets were used to analyse data related to rSECEE accounts:

- a) Reference to examinations: frequency of referring to test-taking techniques; and frequency of predicting future test questions.
- b) Grammar-translation: frequency of using metalanguage; and frequency of translation at word, phrase, clause, and sentence levels.
- c) Focus on form: frequency of teacher's feedback to students' utterances with focus on form; and frequency of explanation of sentence structures.
- d) Aural/oral aspects of English: length of time spent on formal listening practice; frequency of oral practice at word, phrase, clause, and sentence levels; and frequency of utterances made in English to exchange genuine information (e.g., giving instructions, etc.) rather than mechanical oral practice (e.g., reading aloud from the text, etc.).
- e) Request for information by students: frequency of students' asking questions, asking for repetition, etc. (as observable evidence of students' motivation).

- f) Classroom organization patterns: length of time spent on lock-step (i.e., teacher-fronted), pair-work, group-work, oral presentation by students, and individual seat work.

To visually represent the observational data elicited from the 18 classroom observations, each teacher's observed class was illustrated in a graph and compared. The horizontal axis represents the language skills and other classroom activities that were present during the observation period of each teacher, while the vertical axis represents the parameter of time in minutes. As mentioned earlier, the class time for each observation was 45 minutes for each class ( $t = 45$  min), and there were 5 classroom observations ( $n = 5$ ) for each of the three teachers ( $n = 3$ ), thus the total observed time for each teacher is  $t = 225$  min, and the total observed time for the three teachers is  $t = 675$  min. However, the language skills and other classroom activities that were present during the observation period of each teacher were analysed and described within 200 minutes for each teacher because five minutes of every lesson was lost in the opening and closing of the class. The percentage of time spent on each activity was compared for Teachers A, B, and C and is reported in the Results section below.

#### **7.4.3 Semi Structured and Focus Group Interviews with Teachers and Students**

Analysis of the observations led to a number of conclusions, which guided the flow of the two types of interviews. The first step of analysis was transcribing the three teacher interviews. I transcribed the interviews as soon as possible. However, as I only had a limited time at the research site and tried to collect as much data as I could, a basic transcription was sometimes performed when an immediate full word-by-word transcription was not possible. I used a tool recommended by Miles, Huberman, and Saldana (2014) and Saldaña (2013), *analytic memos*, which, as noted by Miles et al. (2014), are mainly conceptual. They not only report data, but also

“tie together different pieces of data into a recognizable cluster, often to show that those data are instances of a general concept” (Miles et al., 2014, p. 30). I was able to transcribe and translate the first teachers’ interviews, which were later checked for accuracy by Teacher HM.

Afterwards, in accordance with Sandelowski (1995), I started a rudimentary analysis. This type of analysis often “begins when the researcher proofs transcripts against the audiotaped interviews from which they were prepared” (Sandelowski, 1995, p. 373). This was the first time I had a sense of the interview as a whole. Despite the fact that I was the one who conducted the interviews and transcribed them, it was at the proofing process that for the first time I actually understood what the participants said. During the proofing process, I underlined key phrases, simply because they made an “inchoate impression” on me (Sandelowski, 1995, p. 373). In order not to lose any line of thinking while proof reading, I inserted comments in the margin next to the text that triggered these comments. Later, I read the interview transcripts several times, to capture the essential features, without “feeling pressured to move forward analytically” (Sandelowski, 1995, p. 373). Each time I read the transcripts new thoughts were prompted about the washback affect of the rSECEE.

The next step for the data analysis was to approach the data systematically by following a data reduction framework (Sandelowski, 1995). Having followed a procedure that guided the generated data in both the teachers’ semi-structured interviews and the student focus group interviews the topics or questions that were used in the interviews served as an organising framework. As recommended by Sandelowski (1995) the data was “segmented according to the responses to each question” or topical area generated by me (p. 375). Building on Sandelowski’s (1993, 1995) rationale, I employed this organising framework in my data analysis reduction stage in order to help put the data in the most usable possible form, and to allow me to see all the

data in a new way. Finally, the interview data was closely examined and thematically coded (Miles et al., 2014; Saldaña, 2013) and themes were compared across the interviews and eventually organised into themes related to the study's research questions. These themes included *washback on teachers* which included accounts on the revised curriculum, and the rSECEE reported impact on feelings, teaching content and practices, learning, classroom testing and students. The second theme was *washback on learners*, which included students' accounts of rSECEE and the Grade 12 EFL instructional and learning practices. Importantly, and as recommended by Miles and Huberman (1994), I attempted to build a logical chain of evidence to achieve an intelligible understanding of the three sets of interview data.

## 7.5. Results

This section reports the findings from the teachers' and students' questionnaires, observations, and semi-structured and focus group interviews in order to address the third research question: *What is the nature and scope of the washback (if any) of the rSECEE on the Libyan EFL Grade 12 teaching and learning?*

### 7.5.1. Washback on Teachers

#### 7.5.1.1 The Revised Curriculum

In the teacher questionnaire and interviews the three teacher participants described the curriculum as good because it can promote opportunities for language learning and language use. Teachers B and C both stated that if the curriculum was properly operationalised in the classroom, students would learn a great deal from it and could experience real language use within the classroom.

Teacher A: Oh definitely, it [referring to the textbook] has been a good change. If I look back at to our days, there has definitely been a huge jump and improvement in the quality

of curriculum. But it is limited with the lack of facilities, and the ministry not providing us with the other supporting resources.

Teacher B: Yes, the curriculum is really good, and very informative. If curriculum is really operated in the classroom with out the influence of the test, it would be very useful for their [students'] actual language use and learning.

Teacher C: The curriculum to some extent is good, in the hands of a creative teacher you can do a lot with it.

However, they reported that the implementation of the revised secondary level curriculum was difficult. In the teacher questionnaire, Teacher A and B highlighted that time pressures, students' attitudes, and classroom norms, were the main stumbling blocks they encountered when implementing the revised curriculum. In addition, Teacher B believed that the current institutional regulations and facilities were impediments. Similar to Teacher A and B, Teacher C considered students' attitudes as a stumbling block impeding the implementation of the revised curriculum. In contrast to the other teachers, Teacher C highlighted that both the prevailing Libyan culture of learning and the lack of facilities at her current school were other influencing factors. The above concerns were reflected in the interviews:

Teacher A: I don't have time to cover everything in the curriculum. I have eight units and each unit has tremendous amount of information to cover within the limited time I have... There isn't enough time to cover each unit thoroughly, there is not enough time to practice or prepare.

Teacher B: The pressure of time, the curriculum which I have to cover for the exam... the majority of the students are demotivated... Plus, there are no facilities that can support you to teach the revised curriculum which in my opinion is quite a good one.

Teacher C: The current prevailing culture of learning and how it is controlling it [curriculum]. The culture of teaching and learning have stopped generations from being actual thinkers.

### 7.5.1.2 Teachers' Feelings

In the interviews Teachers A and B expressed negative attitudes towards the rSECEE, in terms of its format, and the tremendous pressure that it placed on them.

Teacher A: It's a lot of pressure for a Grade 12 teacher. The teacher is put under tremendous amount of pressure you're constantly working for 8 months non-stop. It's very exhausting.

Teacher B: The curriculum that I have to cover for the exam and the time factor all build too much pressure and stress. The curriculum doesn't fully cover the grammar rules, so this requires me to do extra work and search with Mr. Google. The stress I was put through because of the test was intense.

Teacher B also expressed negative sentiments in terms of the rSECEE's limitations which, according to her, include:

1. the underrepresentation of the target construct, where test content is limited to one specific domain of knowledge, i.e., grammar;
2. the failure of the test-score that the rSECEE yields to provide an accurate representation of students' language abilities; and
3. the inappropriate purpose of the rSECEE (given its mandate), i.e. it is an achievement test rather than a proficiency test.

The following quote echoes the Phase I findings:

Teacher B: Especially this year's test<sup>34</sup>, it was just a game. Students spend a whole year learning content knowledge about different topics with all the related vocabulary and grammar rules and then the exam just focusses on just one grammar rule. It limited

---

<sup>34</sup> It should be noted that whenever the participants refer to the rSECEE in their quotes, I use the word 'test'.

everything that could be tested in three pages just to one grammar rule: *reported speech*. The questions were very silly, they chose pointless things to test students on... How test developers just played around with the questions and how the test was unreliable and underrepresenting the whole curricular content covered through out the school year... It's an achievement test not ...definitely not ... a proficiency test.

Teacher B further criticized the rSECEE for not adequately measuring real language skills and use, instead, it was believed by her to be a measurement of memorisation, conscientiousness, and attendance. Teacher A shared similar views:

Teacher A: The test doesn't measure language skills, it measures the ability to memorise and grammar.

Teacher B: Language ability isn't measured here [the rSECEE], but what's measured is students' ability to memorise and the examiners are actually seeing whether the students attended school and classes, in other words attendance, and whether they worked hard through out the year or not.

Importantly, all three teachers believed that the rSECEE is an “unfair” instrument for evaluating Grade 12 students, because in each test there are many misleading items that stop students from answering questions correctly. In addition, test developers are not secondary level subject matter experts, which limits the testing content. In essence, the teachers' beliefs validate the panel review members' and test-developers' responses to the rSECEE, and the Webb model findings reported in Chapter Six.

Teacher A: This is an *evil* exam. it deceives students. The test isn't fair...test developers don't have any knowledge of the curriculum.

Teacher B: It isn't a *fair* test [emphasis added by the teacher]. Limiting the test to one skill isn't fair. Why not test science stream students about a content domain that related to other topics in the EFL textbook, why make it a grammar test?

Teacher B further reflected that the rSECEE was an unfair assessment because students are able to cheat with this type of test. She emphasised that the level of cheating on the revised rSECEE is even more intense than the former rSECEE.

Teacher B: Plus, you need to put a *big thick* line under the word *cheating* [said with a lot of emphasis], cheating on the test is worse these days.

In contrast, Teacher C did not express a negative attitude towards the rSECEE. She reflected that it had not caused stress or anxiety in her classroom. Furthermore, the observational data supported Teacher C's claim, as the instances of laughter within her classroom were nearly ten times those in Teacher A's classes, and three times those in Teacher B's classes.

Teacher C: I don't experience pressure, stress or anxiety during the year. Instead, I am relaxed, and my class is my world... The exam for me and my students has never been a worry or a problem, I don't mention or emphasize its existence in the class.

Nevertheless, she expressed negative attitudes towards the current testing system. She reported that the revised testing system is not compatible with the standardised curriculum; hence, echoing the Phase I findings.

Teacher C: But I have to say what I am angry about is the new testing system. I am totally against the new testing system; discrete item testing isn't in line with our curriculum.

Teachers A and B further conveyed that their students and their success were dependent on them, and that they felt accountable towards students, their parents, and school principal.

Teacher A: It's not just the students who know the test format, but also their parents ... They want happens in the classroom to be similar to the test in content and format.

Teacher A emphasised that the feeling of accountability towards the stakeholders added to their stress. This was echoed even more strongly by Teacher B, who experienced such levels of stress and sense of accountably during the data collection period that her health deteriorated.

Teacher B: This year, the pressure from the test, the feeling of accountability towards the students and their parents, and towards the school really affected my health status. I suffered a lot from low hemoglobin level this year.

Indeed, Teacher A commented that due to the stress and work overload, she decided to opt out from teaching Grade 12 EFL classes for one year.

Teacher A: I feel accountable towards the students and their parents more than the school. It's not just language inspectors are chasing you, but it's also the parents and the school principal. All the eyes are on you...If the students fail the test, the teacher takes the responsibility and she is the one to blame for the failure...The Grade 12 pressure pushed me to take a year off teaching Grade 12 students and to be a supply teacher for a couple of months.

In contrast, Teacher C reported a different form of accountability. She felt accountable towards her students in terms of promoting their communicative competence and developing autonomy.

Teacher C: I push them, I push them to use the language, I push them to take responsibility of their learning I want them to think of English as if we are playing a game. I want them to really have fun and enjoy the experience of learning real English. I motivate them, I won't start my lesson unless they are motivated students, I push and start with the least motivated. In my class, I nurture a human being... I want to help them be good human beings. I am not dealing with puppets... I don't instill information, I promote learning.

In addition, Teacher C felt accountable for promoting: critical thinkers; recognising and appreciating differences; and respecting the learning environment among her Grade 12 students.

Teacher C: When I teach these days, I beg my students to ask me why? Why is it this why? I want my students to be analysers of information...I start to build their confidence, encourage them and thank them for any effort they put. I tell stories about respecting the learning environment. I tell them if you want to learn and be successful in life, it starts with knowing who you are and respecting the people around you, and that you should have your

own identity and not just a copy past of another... Its challenging but at the end we get there together.

Moreover, regarding teacher's roles, Teachers A and B highlighted that their students see them as the "most knowledgeable", "the expert", and "spoon-feeders" of information. Teacher B expressed the view that she was not there to teach the language, but only to prepare students for the rSECEE. Arguably, the teachers' accounts of their teacher role was to "make the ... [student] understand" rather than promote learning (Wall, 2007, p. 147).

Teacher A: Whatever I say in the class they write it down. They know that I know what is best for them and I am here for their success.

Teacher B: I am not happy about my current teaching. Instead of actually teaching a language, what I do is just instill information to students in order for them to pass their test

In contrast, Teacher C saw herself as having a more positive role in the classroom, she described herself as a monitor, facilitator, guider, and listener. As she reported, by taking on these roles she is viewed by her colleagues, school, and students as different.

Teacher C: For most of the time I don't take the floor in my classes. I just guide and monitor while the students take the floor of the class and they are at the centre of learning. The test has no influence on me, I've been labeled as the odd, the abnormal teacher and in some polite situations the *different*. This labelling has also come from students, I always tell them at the beginning of our journey you're unconscious and muddled, confused you don't know what is happening to you or around you in my class. But you will be in a state of consciousness at the end.

It is worth noting that, despite the negative accounts, all three teachers expressed no knowledge of the objectives behind the changes to either the revised testing system in general or the rSECEE in particular. Hence, there may be dissonance between the ideologies of policy-makers and the Grade 12 EFL teachers' beliefs. The following quotes suggest this finding:

Teacher A: I really don't know the objectives behind the reform policy, but what I am sure of is anything and everything can happen in Libya.

Teacher B: I really don't know about the objectives, but I guess it is probably because of the marking, making it easier for them, and helping the disadvantaged students. They no longer have to worry about their hand writing with the new form of testing.

### **7.5.1.3 Impact of the rSECEE on the Content of the Teaching**

In terms of Grade 12 EFL classroom instructional content, Teachers A and B reported that their classes focussed on building the linguistic knowledge of the language and were tailored towards the rSECEE content. Teachers A and B further explained that rSECEE influenced the way they used the assigned Grade 12 textbooks in their classrooms. More specifically, Teacher B reflected that she was advised by her colleagues to focus on the test and never spend time on deepening understanding; in their own words it was a “waste of time”.

Teacher B: I was advised that to teach surface issues not in-depth knowledge, so I started to analyse former exam papers to find out what I could focus on in my classes.

Teachers A and B explained that teaching vocabulary, connotations, proverbs, idiomatic expression and target grammatical structures is particularly important and the fundamental focus of their Grade 12 EFL classrooms. They explained that because the rSECEE concentrates on these areas, their focus is important for developing test-taking abilities. In addition, she emphasized that the more grammatical forms and vocabulary items learners knew or memorised, the better opportunity of passing or excelling the rSECEE. With a guilty conscious, Teacher A reported that she narrowed the curriculum by ignoring the actual teaching of any content domain related to speaking, writing and listening. She further added that the former examination allowed her to teach more language skills.

Teacher A: In the former exam, I used to focus on all the skills but this one [i.e., rSECEE] I just focus on grammar and vocabulary.

Teacher B also reported that her pedagogical attention was on the tested curricular content and other untested content material related to speaking, listening and writing was simply marginalised. However, she did state that whenever she did have the opportunity to implement the non-tested content, she would happily do so.

Teacher B: I use some writing, speaking, and listening activities in my classes but as complementary tasks as they are not tested.

Asked about the motive behind their classroom practice of neglecting the listening, speaking, and writing curricular components of the standardised EFL textbooks, they rationalised that these language skills were never evaluated and did not contribute to the requirements of rSECEE.

Teacher A further explained that listening and speaking have never been the focus of evaluation in either the former or revised version of the rSECEE, and consequently, the skills are not taught at the classroom level.

Teacher A: The test affects and controls what I teach my students. I don't teach listening or speaking any more, I used to at first, but I noticed it's eating up so much quality teaching time, so I no longer teach it... If I did teach speaking I would teach exactly like the exam content. True and false. I just focus on reading content information, grammar book...I focus on grammar and vocabulary. I mainly focus on units 3, 4 and 5, because the three units cover grammar. The whole test just focuses on grammar, even in writing skills they ask about grammar of writing something. so that's what I just teach grammar.

Teacher B: I hardly teach any listening any more, because there is no CD or listening labs. I used to bring everything from home, but I noticed that it's eating so much time and most importantly the skill isn't tested...The final exam just focusses on grammar and vocabulary. So, there is no need for technology or phonetics...The main focus is on grammar and the

reading content in the reading texts. But honestly, I focus more on grammar boxes in the student's book. It's as if it was just a grammar test.

Teacher B added that her grammatical content focus not only consisted of ample explanations of the target structures or grammar rules, but also a considerable number of examples to support the target grammar rule. In her post-observational chats and interview she explained that giving students ample examples of the target grammatical structures was two-fold. First, it would help students acquire the grammar rule more easily and then the students would not need to review it at home. The second reason was for predicting future exam questions.

Teacher B: I give a lot of examples because the more grammar exposure the better then they wouldn't have to go home and practice, because I know most of them won't be working at home. So [laughing] we do everything in the classroom, the learning and practice at the same time. And also, these examples, I hope may come as a possible exam question.

In contrast, Teacher C strongly and repeatedly emphasized in the interview that the content of her teaching was guided by the textbook objectives and aimed at promoting communicative knowledge of the language and its use. In addition, she repeatedly acknowledged in her interview and post observation chats, and it was confirmed by her students during the focus group interviews, that the rSECEE plays no role in guiding her delivered content.

Teacher C: The test doesn't influence what or how I teach. I just teach the curriculum and nothing...I teach language and language use and not the test. I have a curriculum and I teach it in the way that I believe is right, and if my students have mastered only just 80% of the delivered content they will be safe on the exam. I emphasize to them just try to focus about the learning the language and forget about the exam, I tell them "because if you think of the exam your achievement and learning is just going to be short-term and after the exam is over you'll have no English" ... Think of it as a game, or as learning general information.

Teacher C's Student: She doesn't teach us grammar or test content... I always ask her what about the grammar and the test. But she doesn't listen to me and instead she tells me just ignore it and focus in class and continue to participate in class and you'll be fine.

When questioned if she gave more weight to grammar given that it is the focus of the rSECEE, Teacher C explained that grammar had equal weight to the other instructional content.

Teacher C: Grammar is just given like any other language skill in my classes. I don't emphasize it more because of the test, I just teach it according to the curricular objectives.

The teachers' accounts of the washback of rSECEE on the content of their Grade 12 EFL teaching were evident in the observational data. As illustrated in Figure 7.2., a significant part of Teachers A and B's classroom teaching content resembled the rSECEE's focus on grammar and vocabulary. Grammar content alone or in combination with content knowledge of the reading texts and vocabulary, was the most common language skill and content delivered by Teachers A and B. However, this pattern was more obvious in Teacher A's classes, where students spent three of the four observed classes on reported speech. In comparison Teacher B and C spent 82% and 18%, respectively, of their classroom time on grammar and vocabulary.

It can be further seen from Figure 7.2 that Teacher A did not teach either speaking or listening. Although, Teacher B paid some attention to content related to speaking and listening, it was much less than Teacher C. In essence, the content in Teacher C's classes was operationalised in her classes according to the textbook objectives. The speaking and listening skills were operationalised through group presentations, pair-work, listening activities, and group-work speaking activities where each activity correspondingly contributed to around 25%, 13%, and 8% of the observed classroom time.

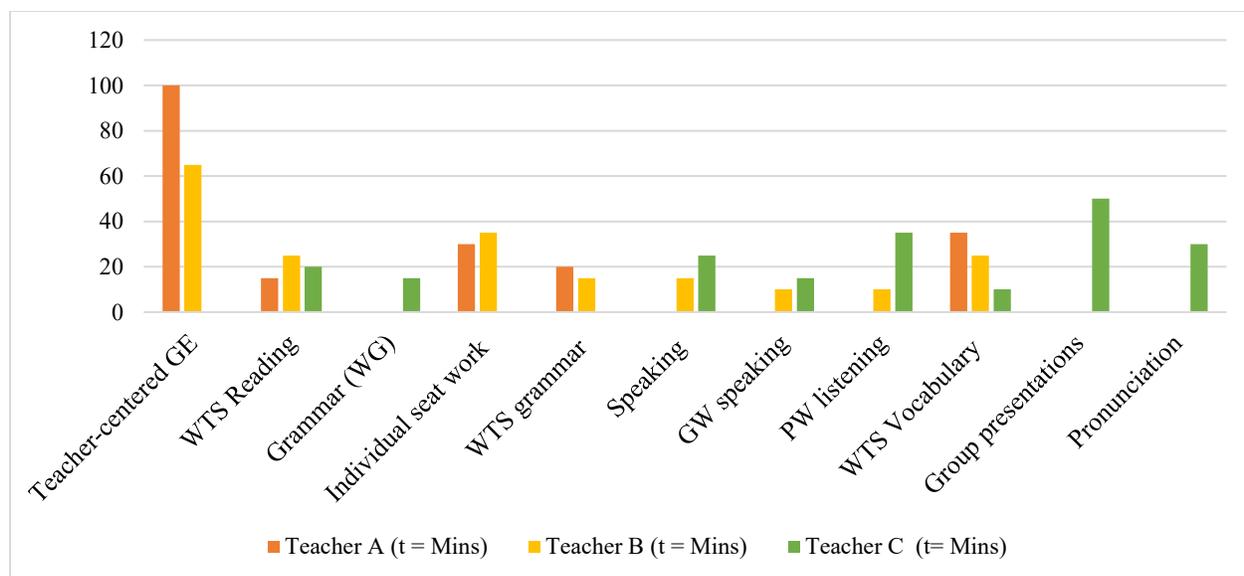


Figure 7.2: Grade 12 EFL Classroom Content and Activities

Note: whole class – involving teacher to students, (WTS), whole class student to students (WSS) Pair work or group work (PW, GW), Individual work (IW), Grammar Explanation (GE).

In addition, despite the fact there is no oral assessment on the final exam, Teacher C focused on the pronunciation and transcription of words, emphasized sound differences throughout the observations. These contributed 15% of the observations time. Explaining her focus on pronunciation and homonyms she stated that every component of language is important if the aim is to teach language and to promote language use.

Teacher C: If you want to be special, you have to focus on the details, like pronunciation, intonation.

#### 7.5.1.4 Impact of rSECEE on Methods of Teaching.

Teachers A and B emphasized in the post-observation chats and interviews that their Grade 12 instructional pedagogy was affected by the rSECEE, which encouraged them to take approaches they would not do otherwise. Teacher A explained that she as a teacher had changed as a result of implementing the revised testing system in the Libyan secondary level.

Teacher A: I as a teacher have changed after the rSECEE and the multiple-choice questions...Complaints from parents, complaints from inspectors about they way I teach and after having the new test in place. They [refereeing to language inspectors] said that the test question are multiple choice questions, so and your teaching and testing in the classroom should reflect this.

Furthermore, Teacher B noted that she was only teaching students language rules and vocabulary, which in her own words did not represent “real English”.

Teacher B: The existence of the test dictates to me from the first day of the school, what I teach and how I teach. I have to cover the test curricula content without being innovative and creative. The test drives everything that I do and how I do it in my classroom. when I start preparing my lesson, I look at the textbook, I highlight what is important in terms of test not in terms of teaching the language...For every lesson, the important information, vocabulary content is identified, then I decide how I will deliver it to the student in line with the test’s requirements.

Teacher B used an analogy to reflect on the intensity of the rSECEE’s impact on her Grade 12 EFL classrooms. She described the power of testing as a man giving out strict orders, and therefore she had to sacrifice her beliefs about language pedagogy to obey the orders.

Teacher B: Let me sum it for you, forgetting about my beliefs about teaching and learning, you have my communicative language approaches to teaching English which includes the proper teaching of reading, pronunciation, group-work, pair-work and my English language teaching resources on a table and then you flip over the table with everything on it. After everything falls off the table, comes this man with all the confidence in the world and slams down the test and says “This is what it is!” and then he slams down the textbooks and walks away. In a nutshell, Grade 12 EFL teaching is just the test, and unfortunately nothing more.

Teacher B further added that in absence of the test, she would be more innovative and employ communicative approaches to EFL classroom teaching and assessment. The employed material would be more authentic and actually reflect authentic English language usage.

Teacher B: In a scenario where the test didn't exist, I would be living in a fantasy world of English teaching. I would be really enjoying my job. I would be innovative, I would be actually doing real language teaching with a lot of emphasis on productive skills and especially the students' oral skills. If there was no test and I was responsible for assessing my students, I wouldn't have a high stakes test and I would implement formative assessment tools. ... But at the end of day I am not happy with what I am doing, because what I am teaching isn't real English.

Teacher B emphasised strongly in the interview that teaching Grade 12 EFL classes was like teaching test preparation classes. She argued that teaching has to be effective in terms of helping students to pass the test so that students' is not jeopardized by the rSECEE.

Teacher B: For sure the rSECEE influences how I teach, when you see the questions at of the end of the year and how students are jeopardised with such questions you start to reflect and say to yourself "I won't let this happen to my students and I will do what is in their best interest to help them pass the test". So, what I do is, I prepare them very well for the test, just like an exam preparation course.

In Teachers A and B's classrooms, Arabic was used more as the medium of instruction and communication, whereas the use of English terms or phrases was emphasized when the teachers were explaining English grammar rules; for example, okay listen to me, open books, verb, noun, statement, wh. questions, past tense and so on. Teachers A and B explained that they used both languages to cater for the different needs and abilities in their classes. In addition, they stated that ministry officials, in particular language inspectors, emphasised that both English and Arabic should be employed in the Grade 12 EFL classroom, especially when the focus of the lesson is

grammar. This was justified because the students' language competency levels did not permit for English to be the only medium of instruction.

Teacher B: I have to use Arabic and English to manage the needs of all students. there is a variety of levels in the classroom and I have to attend to those varying needs. Plus, an inspector came once to evaluate my EFL classroom practices an he told me to use more Arabic and less English, especially when I explain grammatical structures, he said “you need to make the matter easier for them and not complicate it. They have other subjects than English to deal with”.

However, Teacher C reported in the interview that the rSECEE had no affect on how she delivered the standardised EFL textbook. As a result, a high proportion of the language instruction and communication (approximately 70%) within Teacher C's Grade 12 classes was in English.

Teacher C: I have the course objective and the textbook, why would I need the high-stakes test at my table. The test doesn't drive my students or me; instead we're the ones that drive the test.

It is important to acknowledge that Teacher C recognised the impact that rSECEE may have on many Libyan teachers EFL instruction and she emphasised that teaching to the test is the norm for most Grade 12 classes in Libya. With such awareness she has come to categorise herself as “different”.

Teacher C: The convention for Grade 12 teaching is to teach to the test and only to test. The teacher will emphasize the exam and ask students to highlight only things that will be tested. But for me it's a totally different case. I don't teach this way, because I am dealing with wave of generations one year after the other and they will be entering university where the medium of instruction is mostly English, so language use is a must.

All the teachers revealed in their questionnaire answers that they were proponents of communicative language teaching approaches. Teachers A and B considered this approach to be

a combination of teacher and student-centered approach, while Teacher C considered her classroom practices to be oriented towards student-centered learning. However, in contrast to their questionnaire answers, Teachers A and B employed a grammar translation method (White, 1989). for teaching grammar and vocabulary. When questioned about this anomaly, both teachers stated that they did not employ the communicative language teaching approach because it was incompatible with the principles underlying the rSECEE and the method did not meet with the students' requested needs.

Teacher A: I used to teach all language skills when we had open-ended response questions.

Teacher B: As a teacher you want to use PowerPoint, you want to be innovative in teaching to try out approaches. But you come to the class and find the students saying, "we don't have time for all of this, just give us the important notes show us what to underline and memorise and that's it".

Furthermore, Teacher B argued that top-down policy makers dictate what teachers do in the classroom. She stated that the language inspectors, who visit twice every year to evaluate EFL classroom teaching, advised her to stop her communicative practices and employ the conventional Libyan approach to language teaching (see Sections 2.4.2 and 2.4.4) instead. Consequently, and unwillingly, she gave up her communicative approaches at the expense of satisfying stakeholders.

Teacher B: They told me that you'll start from beginning as you'll have go back to your old methodology. They asked me to not put too much effort in my teaching and make my classroom tests as easy and approachable as possible for the students.

The analysis of the observational data also reflected the discrepancy in Teachers A and B's questionnaire answers in relation as to who was mainly responsible for controlling the learning within the classroom. The discrepancy demonstrates the inconsistency between their BAK

(Woods, 1996) and their real actions in the classroom; thus, what the teachers think they doing does not match what they are actually doing in the classroom (Woods, 1996). The main discrepancy found was in terms who was controlling the flow of the lesson. As seen in Figure 7.2, Teacher A was the centre of focus of her classes; nearly 85% of her classroom teaching time was teacher-centered, with a degree of student interaction in certain activities. In the case of Teacher B, approximately 65% of her classroom practices were teacher-centered with student answering and requesting information. In contrast, Teacher C had no teacher-centered activities that focussed on the teaching of grammatical structures.

It was further observed that Teacher B was using the grammar translation method of language teaching less frequently than Teacher A. This observational finding is clearly shown in Figure 7.3, with Teachers A and B having approximately 60% and 50% of their classroom activities mirroring the grammar-translation teaching approach (see Section 2.4 for definition).

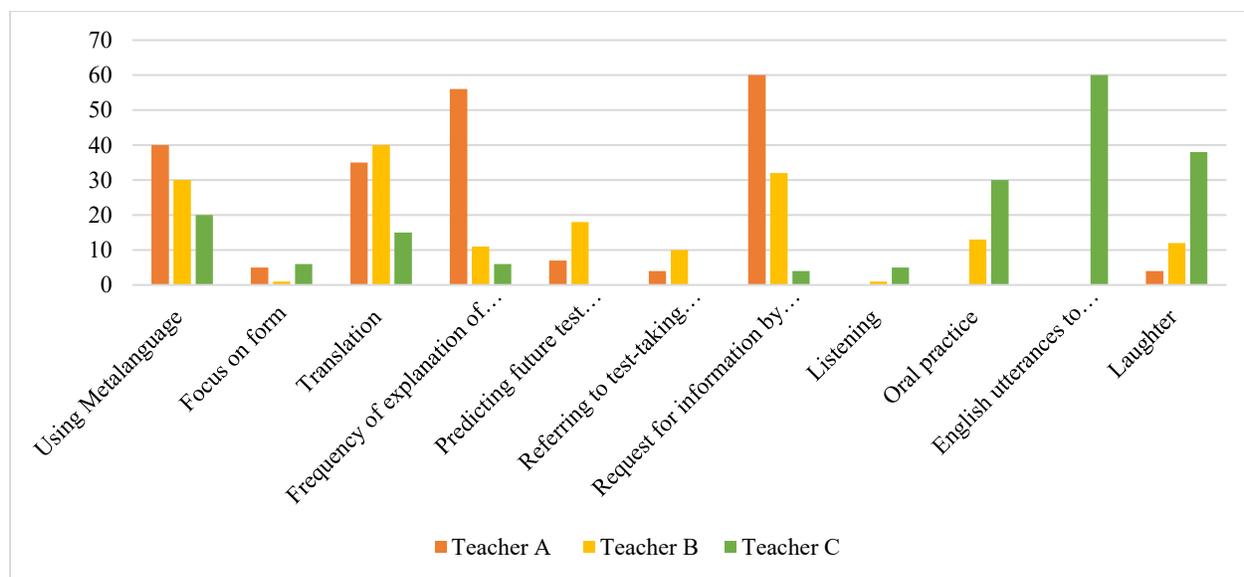


Figure 7.3: Frequency of Classroom Activities of the Three Grade 12 EFL Classrooms

Note: SS = sentence structure

Teacher C had a portion of her classroom activities reflecting the grammar-translation approach of teaching. However, with only 18% of her employed teaching activities being

oriented towards the grammar-translation, it can be argued that it was minimal in comparison to Teachers A and B. Both teachers said that they were using the grammar translation method because “it was the most appropriate method to cover the curriculum and meet the test demands within the time frame we have”. They further reported that this method allows them to cover the test content and format with greater thoroughness.

Teachers A and B’s classes drew on the rSECEE’s restricted format and presentation of information. Information related to grammar and vocabulary was presented in a decontextualized manner. It was observed that Teacher A would write down the grammar rule on the board and then explain it in Arabic. The students would copy everything that she had written on the board. After a couple of examples and with minimal student interaction she would ask the students to open their workbooks and then start to elicit answers from the students. On many occasions students did not interact with her and she provided the answers herself while the students recorded them.

Teacher A: This is my way to teach [referring to decontextualized teaching of grammar], you have to teach grammar on the board. Students need to have the rules in Arabic and the inspectors told me to do so.

In contrast, Teacher B wrote a number of examples representing the target grammatical structure and explained the rule through the use of examples. She would then ask students questions in English and help students understand the grammar rule, but without giving them sufficient time to answer. There was greater student involvement in Teacher B’s classes than in Teacher A’s, where the only teacher-student interaction that took place was when students requested information or clarification of word or grammatical structures. Teacher A’s students requested information or clarification 15 times more than did Teacher C’s students and approximately twice as often as Teacher B’s students (see Figure 7.3). Importantly, Teacher C

taught grammar differently by integrating it with other language skills and students worked in groups to deduce the target grammar rule. This activity took only around 8% of her observed classes, which is an indicator of grammar not being the focus of the teaching. In other words, it is an indicator of the rSECEE of not having a negative washback on either the what and how of Teacher C's teaching.

Content related to the reading component of the textbook was also taught in a decontextualized manner resembling the rSECEE format by Teacher A. It was further observed from both the classroom observations and post-observation chats that she never allowed her students to read. Instead, she would read the text three times and then provide a literal translation of the text. The time, accountability towards covering the whole curriculum, and students' poor reading abilities were factors that Teacher A chose to justify her grammar-based approach for teaching reading.

Teacher A: I don't let my students read aloud in my classes, and some students complain about this. Students reading in the classroom means wasting time, and I don't have time to waste. I read the text very clearly three times, work on the vocabulary and give the students answers to the reading exercise. Students are weak at reading, and if they read it means I have to correct, and I don't have the time, I need to finish the curriculum. Of course, they [students] complain about not allowing them to read whilst their former teachers did. I tell them "you're allowed to read in grade 10 and 11 because you didn't have a high-stakes test at the end. But in Grade 12 you do, and I don't have any time to waste". If I give them a chance to read that means, I'll have to spend five lessons on one reading text.

Teacher B: As for the reading text that we cover in class, I take this text and make my own questions about this text. Because the test asks question in between the lines and I know how they test these kinds of things. This question can be multiple choice, or true and false.

Perhaps because Teacher B's belief system was more oriented towards the communicative teaching approach, she allowed students to read. Indeed, she asked students to show emotion by

adding intonation to their reading, as well as focussing a couple of times on pronunciation. She also asked students to bring in artefacts (e.g., a family member's vaccination card) in order to "get away from memorisation of facts and learning of facts". She explained that getting students involved when the focus is on the reading content of the curriculum would help them become active learners.

Teacher B: I implement every now and then some activities [such as read aloud] ...to encourage students to be more active learners and change the classroom atmosphere and break the tedious routine on grammatical practice.

As with the teaching of grammar, Teacher C took a different perspective in teaching reading and its related content. It was observed that she spent a considerable amount of time on developing students' understanding of the topic, expanding the reading topic, and highlighting the different word formats of the target vocabulary. For example, one of her observed classes was on the topic of motivation. Teacher C operationalised the concept of motivation in her class with real life examples so that students could understand the English connotations related to the word. In addition, she asked the questions in English and did not provide a direct translation of the target vocabulary. Instead, she encouraged her students to speak in English and provide a definition of the target vocabulary in their own words. The following quote illustrates this approach:

Teacher C: I have to make them active by making boring topics or units in the textbook to be more interesting and get my students more involved and honestly this can be very challenging at times.

Another classroom activity that mirrored the grammar translation method observed in both Teacher A and B's Grade 12 classes was the use of individual student seatwork. In the cases of individual student seatwork for both teachers, students were either individually attempting to

answer course/workbook exercises related to grammar, vocabulary, or reading comprehension questions. As highlighted in Figure 7.2, Teachers A and B had around 15% and 18% respectively of their classroom instructional practices devoted to individual student seatwork. By contrast, Teacher C had no individual student seatwork in her classroom time, with this time being oriented towards more student-centered activities. This finding related to Teacher C's instructional practices was opposed to the one predicted in Chapter Five, and, in turn, could be interpreted as an indication of positive washback. Although there was limited practice of the listening skill in Teacher C's classes, at approximately 18% of classroom time, she implemented student presentations in nearly all of five observed classes and encouraged the non-Grade 12 student presenters to actively listen to their peers' presentations because they had to take notes. Teacher C was also frequently exchanged information with her students in English. This frequency was very high compared to Teachers A and B.

The frequency of oral tasks in Teacher C's classes was also very high in comparison to Teachers A and B. Furthermore, in comparison to Teachers A and B the practice was not mechanical, such as by reading aloud from textbook or repeating after the teacher, rather Teacher C's students were actively involved in delivering presentations in their own words. It is important to note that Teacher C's classes were more innovative (see Glossary of Terms) compared to the other two teachers. The classroom atmosphere was lively and energetic, which may have positively influenced her students' motivation and attitudes towards learning English.

Teacher C: I have to tell you something from the beginning, I am the type of teacher who won't be happy if my students aren't active with me in the classroom, them taking a passive role in the classroom inevitably effects my teaching too. It can have constraint on how and what I give.

Overall, Teacher C and her students used a range of skills and covered all language skills equally. Her classes were student-centered, and she spent considerably more time on classroom activities that involved student interaction, speaking and writing than did either Teacher A or B. The following quote that was articulated in English illustrates the above finding.

I just want to make English as fun as I can, and I want them [students] to use English in their lives and I make it very interesting, I always tell them it is as if we are playing a game...I say to my students learning English is yummy as yummy as a chocolate cake. Don't complicate everything just listen to me and you will learn.

Both Teacher A and B highlighted the importance of exam preparation, which consequently meant their main common concern was to fully prepare the students for the high-stakes exam, in the limited time available. Both teachers stated that they employed test-taking practices. According to their reflections, it was inevitable because of the stakes involved. They further added this approach was expected.

Teacher A: It is a given to prepare your students for the test.

Similarly, Teacher B reported spending reasonable time preparing her students. Both Teacher A and B reported that they started exam preparation towards mid-May, but they repeatedly emphasized that their teaching from the start of the school year was actually preparing the students for the high-stakes test.

Teacher A: I constantly say through out the year "You need to underline this and that, memorise these words, make sure you know them, they are important".

Teacher B: Throughout the year I tell my students to underline this and that, and I actually go around the class to check that each student has actually underlined it.

They considered the last six weeks of school a period in which the teacher would clarify or review problematic areas raised by their students. The rSECEE influenced the types of material the teacher used during the exam preparation period. Both teachers used past papers and

commercial publications that contained model rSECEE questions. They further emphasized that in a two-week period before the end of the term, they gave daily test practices and mock examinations. They added that they used their own supplementary material, which they developed for their students. They explained that they kept a record of all the possible questions that their students could encounter in the rSECEE. Interestingly, Teacher B compared her test preparation process to building a boat, in which she expected her students to sail on the test day.

Teacher B: I go over exam papers cover each and every question, use commercial books...I built the boat thorough out the year with fully preparing them for the test and on the day of the test, I tell my girls you're the ones who'll sail the boat.

Both teachers also emphasized that not only students would be upset if the test preparation practices did not serve their needs, but also that their parents and guardians would be unhappy and would even lodge a complaint against them.

Teacher A: Every week, I give my students tests, each week testing a different test technique one week it will be just multiple choice the other it will be MCQs and so on. With this preparation that I do on a weekly basis, the students are well trained and prepared ... I review everything before the exam, go over past exam papers and have mock tests.

In contrast to her colleagues, Teacher C reported that she never employed test preparation practices in her classroom. Instead, she noted that she would make jokes about commercial publications that consisted of model rSECEE questions and asked her students never to refer to them. She also reported that she had constantly emphasized that if her students focussed and learnt the language and the content delivered to them by either herself or their classmates it would be more than enough to excel in the final exam.

You asked me about test preparation, this notion doesn't exist in my class... I don't use those commercial books and I tell my students if you want to pass this exam and be good speaker and writer of English don't even go near Ms. Samaria's recipe book.

Teacher C, however, did emphasize a reoccurring situation in her classroom, which Teacher B also raised. Her students regularly asked about important content information on which they needed to focus. She said that she would provide them with an unexpected answer: “everything is important”. She calmed students’ nerves by reassuring them that as soon as they learnt how to use the language effectively, their instinct would start to work, and they could then deduce what was important.

I tell my students what's important or not? For me everything is important. I tell them you can judge which content from your understanding of the topic what sounds important and deserve extra work.

In addition, there were differences in the way teachers emphasized curricular content in relation to the possible rSECEE content. As a percentage of the total class time, students of Teacher A received three times as much information about the rSECEE than those of Teacher B, and 30 times much as Teacher C’s students. This was a possible rationale for why Teachers A and B were the centre of the classroom activities.

#### **7.5.1.5 Impact of rSECEE on Learning.**

Teachers A and B reported that the revised testing system has marginalised learning. Teacher A argued that the actual learning that took place in Grade EFL classes is restricted to the enhancement of the ability to memorise and recall, thus echoing the Phase I finding. According to Teacher A, no deep learning takes place because students have difficulty with understanding the question stems of the rSECEE. She explained quoting her students words, that students were able to successfully answer many test questions by just memorising parts of the question.

Teacher A: Students have difficulty even understanding the stem of the items... the only skills that students learnt to use is memorization and recall...students have said even if we memorise part of the question we can still get the answer correct in the test.

Similarly, Teacher B reported that the existence of the high stakes testing system within the Libyan educational context has influenced English language learning of her students and the majority of Libyan students. She blames the rSECEE for making the learning of English in the form of discrete elements, and the systematic rote learning of facts and information; in her words “unproductive learning”. She also believed that the rSECEE created a lack of students’ autonomy and added to the students’ inability to communicate in English. Very emotionally, she noted that although her students had sound linguistic competence (such as knowledge of the grammatical system and vocabulary), they would not be able to put this knowledge into practice when faced with real life situations. Her reflections were clearly mirrored in her class when she encouraged students to communicate the information in English, but the students were unable to accumulate or systematize the information that they had acquired from the teaching practices relating either to grammar and vocabulary.

Teacher A: No real or actual learning that takes place during the Grade 12 academic year, even though we have a really good curriculum but after the exam is over the student has actually learnt nothing.

Teacher B: The exam has an affect on the students’ learning, the students aren’t actually learning, they learn what is delivered to them in class for the sake of passing the exam, they are learning for an end product and goal which is to enter university ... Because of the way we teach English which as I said before is in accordance with the test requirements, students can not communicate, students are no longer able to write a short composition. Students no longer pay attention to spelling, learning is basically for passing the test only.

To help overcome the limitations of the rSECEE in her classroom, Teacher B stated that she tries her best to provide constructive feedback during the limited pair-work activities, as well as to encourage her students to be active learners and asks them to figure out the answers themselves. This practice was confirmed during the observations. In addition, in the interview

she stressed that language use is not promoted in the classroom, hence highlighting another negative washback of the rSECEE.

Teacher B: I go over the curriculum to highlight important components to cover in terms of test not in terms of the promoting students' language use or language learning, we don't have that kind of learning here... I try to provide feedback on pair-work. I implement every know and then active student reading activities or pair work to encourage students to be more active learners and change the classroom atmosphere.

Teacher C reflected on how the revised testing system impacted learning within the wider Libyan educational context, rather than how the rSECEE was impacting learning in her language classroom. She argued that learning in schools could be described as the transformation of knowledge, which has been aggravated following the implementation of the revised testing system.

Teacher C: They have put English in boxes, and language in a box full of chances, where students can just primarily guess and pass.

The shortcomings of the revised testing system in Libya, as reported by Teacher C, has marginalised students' level of thinking, whereby students are unable to produce or analyse information, and continue to be passive learners.

Teacher C: The whole thing is wrong, students are no longer thinkers since the implementation of this form of testing. Our students no longer have the motivation to learn or further their knowledge in different areas of interest. We no longer have thinkers or analysers in our schools ... Because of our imbedded testing system, we no longer have open minded and reflective students. The students are still receivers of information and the teachers are the spoon-feeders of information.

Although C reported that the new testing system in general was marginalising learning within the education system, she believed that real learning was taking place in her current Grade 12 classrooms. She explained her approach is always challenging, but she is happy and feels her

efforts are worthwhile when students communicate and put language into practice. She further highlighted the importance of the relationship between the students' sense of belonging and learning. She reported that she made students feel that they were all important to her.

Teacher C: I put in all of my efforts and it is very challenging, but at the end of day, I'm very happy with the end product, they have actually learnt something, and true learning does change people ... I strongly emphasize to my students that every student in this class is important to me and I am not here just to give information and leave. I am not a spoon-feeder, I am a facilitator.

Her statement was confirmed during the classroom observation, where students were: using English to deliver information in their group presentation; providing their own definition of words in English; and requesting information in English. As an observer, I was surprised to hear students asking to be excused to go to the bathroom in English. The structures of the sentences were grammatically correct, and the students' application of skills to real-world problems were not a "separate and difficult learning hurdle" (Shepard, 1991, p. 233). The following quote was articulated in English by a student in one of the observed classes:

Student of Teacher C: Teacher, could I please go to the toilet [British word for bathroom]?

#### **7.5.1.6 Impact of rSECEE on Classroom Assessment**

From the interviews and the copies of the in-house classroom assessment (see Appendix S) that teachers brought with them to the interviews, there was washback from the rSECEE on the three teachers' classroom assessments in terms of their content and format. The three teachers further reported that they relied on the textbook content for their in-house classroom assessments, which, in turn, mediated the washback of the rSECEE.

From Teacher A and B's accounts of the impact of the rSECEE on their classroom practices, it was clear that the high-stakes test affected the way they evaluated their students in their

classrooms. They reported that the classroom mirrored the rSECEE even before the mandated top-down policy was issued. The Libyan Ministry of Education mandated that all in-house Grade 12 EFL classroom assessments have to mirror the rSECEE in terms of format and content. In addition, Teachers A and B emphasised that they were influenced prior to the ministry's mandate by the students' parents who wanted the Grade 12 classroom assessment practices to mirror the rSECEE format and content. The teachers reported that the students' parents argued that classroom assessment mirroring the rSECEE would help familiarise their children with the test format, and, thus, prepare them for the rSECEE.

Teacher A: Because of its stakes, I have come to have those evil type of tests in my classrooms. But also, students' parents have put forward plenty of complaints to the principal wanting our tests to be a copy-paste of the final examination.

Teacher B: There was a regulation imposed upon from the ministry to no longer ask students any open response questions, we just ask MCQ questions

Commenting on the topic of test preparation practices during the interview, Teacher A proudly described an anecdote that recalled what happened on the day of the rSECEE. The quote highlights the washback *intensity* of the rSECEE on Teacher A's instructional practices:

Teacher A: On the test day, I went to see my girls [students] during the test and asked them "Girls how is it" they replied that *Oh teacher* [emphasis added] this is your test, it as if I had put the test, i.e., from my testing preparation practices. Then I asked them is there anything in this test I didn't give or explain, is there any grammar rule in that test that I didn't thoroughly explain...they all replied *oh no teacher* [emphasis added].

In her comments, Teacher C stated that although her classroom practices do not adhere to either the how or the what of rSECEE, her classroom assessment practices did. She further stated that she was obliged to follow the current ministry mandate regarding the format of the Grade 12 classroom assessment. Nevertheless, she felt she still implemented her philosophy of language

teaching and learning for evaluating the content material of the Grade 12 EFL curriculum. She also emphasised that she assigned students different writing assignments, such as writing reports and book reviews, in order to evaluate her students' application of the target grammatical structures.

Teacher C: This is something that I had no control over, it has been a mandate so playing things safe, I had to do it... My classroom tests are only similar in format but *never* in content I get my students to write reports and book reviews in English. With these different assignments they will be able to demonstrate to me their knowledge of grammar and its application in authentic situations.

#### **7.5.1.7. Impact of rSECEE on Students**

Teachers A and B both commented that the rSECEE had a negative washback on their students' motivation levels which, as they reported, was caused by the pressure the rSECEE generated.

Teacher B: Of course, the exam has an affect on the students, most students aren't actually learning, they learn what is delivered to them in class and for the sake of passing the exam, they are learning for an end product and goal which is to enter university.

Teachers B further explained that the exam and its neglect of one receptive skill (listening) and two productive skills (speaking and writing) has had an impact on students' attitudes towards both the teaching and learning of English. She reported that students were complaining and mainly negative about her initial communicative language approach to Grade 12 EFL teaching, because this did not relate to the rSECEE. She believed that such negative attitude stemmed from students' conceptions about the activities and the status of the rSECEE.

She further reported that her current and former students regularly wanted her to provide them with model exam items and constantly highlight the curricular content that required extra attention.

Teacher B: They are scared of the test, they are constantly asking “will this come on the test, what about this point?” ... if I don't say that this is important and it may be a possible question on the test they will never learn it or pay attention to it.

In a nutshell, she believed that her students did not enjoy the experience of learning a language, instead they saw the EFL classes as a means of passing the high stakes exam. She explained that their complete focus on the rSECEE and their negative attitudes towards different approaches to learning was justified because they had been reared in the prevailing learning culture over the 12 years of schooling to think this way.

Teacher B: I should have more communicative activities, but the student won't allow me or even accept it at any level ... “the test and nothing more” they would say.

Although Teacher B's students restricted their learning to the test format and content, their teacher claimed they were happy about her Grade 12 EFL classes, as she fulfilled their needs. However, upon analysing the focus group data there were discrepancies between the two stakeholders' accounts.

Teacher B: My students are happy because I'm the one who reads translates explains, and they want to answer the questions too...They are happy because I give them what they are expecting of me, what will be covered on the test, so they are happy!

Student: I don't want a textbook, I want language and how to use language. That's all I want.

Teachers A and B reported that their students became more motivated whenever the class focus was on grammar. This was backed up during the observations of both classes. The students would focus intensely during the teacher-centered grammar explanation activity. They would request information and clarifications in Arabic during this activity. Teacher A's students requested approximately two times (60 times) more information and clarifications than Teacher B's students (32 times).

Teacher A: There is a lot of interaction and students become more motivated with me in grammar lesson because they know that grammar is important for the exam and they need to learn it.

Teacher C provided a different scenario when the topic of students' attitudes was brought to her attention. She explained that her current and former students' attitudes changed from the middle of the year when they became more interested in learning the language and were delighted to have experienced the true meaning of learning a language. She also reported that they become more motivated and had a stronger motive to work harder.

Teacher C: Its challenging but at the end we get there together.

From the analysis of the observation data and coding matrixes, it can be argued that the account was reflected in Teacher C's EFL classrooms. The students were studying in a comfortable relaxing environment because the occasions of laughter (38 times) gave a "general indication of the atmosphere in the classes" (Hayes and Read, 2004, p. 106).

Teacher C: There is a lot of push and pull in my lesson... I try to make my classes interactive and the students take the floor of the conversation...they have the ball. If I was the center of the focus in the classroom, what will they do in the class... I give myself extra work but at the end my students actually learn.

Before, I move on to discuss the students' findings, I would like to end this section with a quote articulated by Teacher C in English, which clearly defines her BAK (Woods, 1996) regarding the rSECEE and how the washback of a high-stake test is operating within her Grade 12 classroom. A more detailed discussion of this account is provided in the discussion section (see Section 7.6).

Teacher C: There is a *gap* between the test, the teacher and the curriculum used in the classroom. And what the teachers does is link the curriculum to the test in the best way possible. So, the teacher is the linkage between the two (Linking her fingers together).

## 7.5.2 Washback on Students

### 7.5.2.1 Students' Attitudes Towards the rSECEE

The results of the 15 students' questionnaire responses are summarised in Figure 7.4. The data highlight that the majority of students: did enjoy learning English (60%), particularly speaking (100%) and listening (73%); expect teachers to prepare for the test (60%); and do test preparation practises outside the classroom (60%). Fifty three percent (8 students) of the sample reported in the questionnaire that they were worried about the upcoming rSECEE. Although, Teachers A and B believed that students do not like learning English and they are just learning it for the sake of the rSECEE, the student sample reported they enjoyed speaking in English, and 11 (of the 15) students enjoyed listening, indicating discrepancies between the stake-holders' accounts.

It is worth noting that one in three students who stated they were worried about the high-stakes test were Teacher C's students. In addition, four out of the nine students who expected teachers to prepare them for the rSECEE were Teacher C's students. The above finding may suggest that the two key stakeholders (students of Teacher C and Teacher C) shared different views. This difference may be an important factor in understanding the washback of rSECEE on Grade 12 EFL teaching and learning.

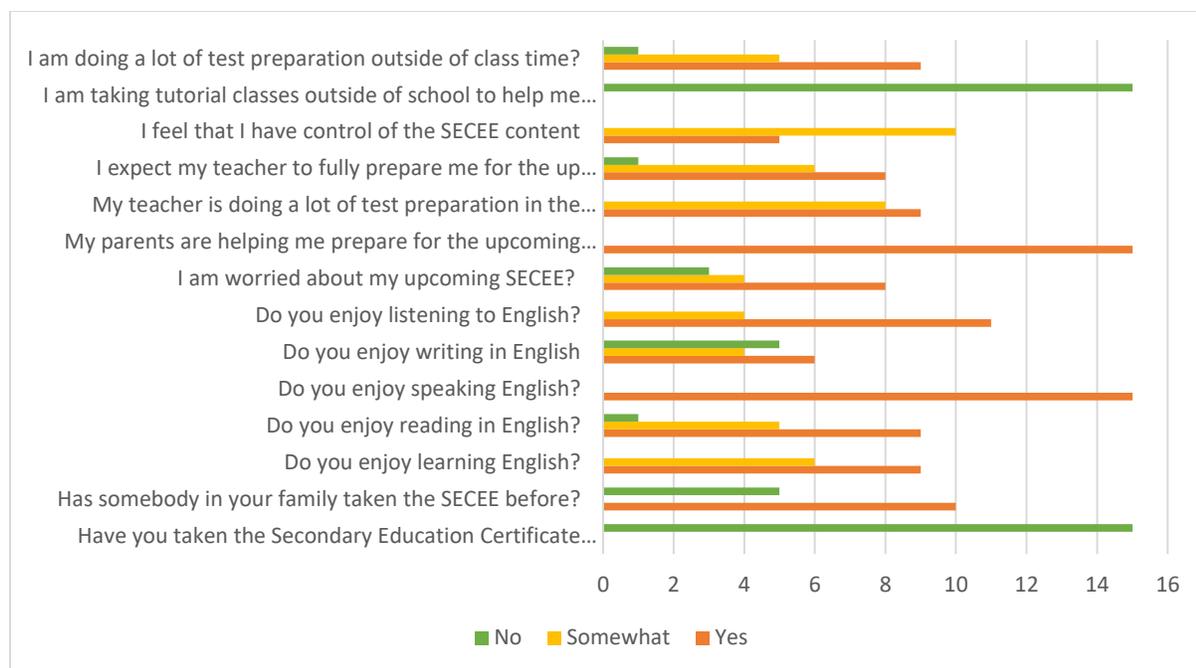


Figure 7.4: Grade 12 Students' BAK towards the rSECEE

The majority of the Grade 12 students reported negative attitudes, scepticism and mistrust in the rSECEE as not being a true representation of their actual language competency.

Student: I think it [rSECEE] reflects hard work and conscientiousness. I don't think it is true representative of true language skills. If I was put in an English language speaking country, I wouldn't be able to communicate with anyone and this is because of the way we're taught English in Libya. With no focus at all on speaking or listening and writing. So, this would be the consequence.

Student: I don't think its important because its just an evaluation of what I know from the curriculum. Its not an evaluator of how well I know English. If I sit and talk with a native speaker, I am not going to say myself oh yeah this something I learnt in Grade 10, 11, or 12 pages so and so.

Student: I like English and can communicate in English very well, but I don't get a good test score sometimes and I have failed my mid-term tests... It [rSECEE] doesn't represent my comprehension level... It [a high rSECEE test score] just says I learnt grammar and vocabulary and I'm a hard worker that's all.

Student: What it reflects is just my ability to memorise, whether I attended the classes and whether I have the ability to apply grammar rules.

Furthermore, Grade 12 students reported negative attitudes also covered the rSECEE having misleading test items, limiting test content to just one domain of knowledge (i.e. grammar), it being unfair, and how guessing can play important role in students' successes on the rSECEE. This is a stakeholder's validation of the alignment model's findings reported in Chapter Seven.

Student: There were many items that actually put you in corner...you don't know what is testing content or grammar.

Student: Its [the rSECEE] not fair, the test doesn't cover the curriculum.

Student: These misleading questions makes you doubt any answer you give, because you think that either way its wrong, in the end they push you to guess.

Student: They just squeezed us in one area: *grammar* and that's it.

Student: The test just tests grammar and sometimes vocabulary ... but just testing grammar isn't fair, we have studied other things.

Moreover, the majority of students reported from the started of the academic year that the rSECEE provoked instances of emotional, psychological and physical distress, which included the effects of confusion, frustration, considerable amount of stress, and anxiety. The closer to the rSECEE, the more stressed and anxious the students reported they become.

Student: From the beginning of the year, I have been in a bubble of stress. This bubble was made by the school, teachers and society but not my parents, who constantly emphasize the importance of Grade 12 tests. Even if I go out with family, people would ask me what year are you in? and I would say Grade 12, they would say "Oh may Allah<sup>35</sup> be with you".

---

<sup>35</sup> The word *Allah*, according to the religion of Islam religion is God

Student: We are always worried, about the test. Whenever students met outside of school they would just talk about the Grade 12 tests, and the possibility of failing them.

Student: I even more worried now, these are really stressful days [a week before the test date]. And what's making me even more stressed is the amount of content we are given and the expectations to memorise all information in the limited time we have. It needs a lot of hard work and time.

In addition, the students strongly emphasised how a single test performance decides their future and can stop them entering their desired university programmes.

Student: Of course, except those who have a very good background in English. Most students are worried about the rSECEE, because it carries a large weight in determining which faculty you can enter.

Student: Because I wanted to go into engineering I need to have a high-score on the rSECEE, and I don't want another programme, thinking about all of this made me even more stressed and angry.

Student: I am not worried about passing or failing, what I am worried about most is my test score, will I get low or high one, because either way it will affect my overall average and especially I need a score in the high eighties or nineties to be able to go into medicine.

They reported that because of the anxiety level on the day of the rSECEE, they performed below their actual ability. They reported sadly that some of their friends and classmates experienced physical distress and were crying and having a nervous breakdown during the test, which, in turn, had an affect on their own test performance.

Student: Inside the exam, I was really scared, I was so stressed out and felt frozen and couldn't answer at the beginning because the exam was extremely difficult.

Student: The whole testing room was full of anxiety...students were crying and asking for water because they were unable to answer.

Student: Complete shock, I just kept looking at the girls around me and everyone was just shocked and stressed ... The test was extremely difficult, and I wanted to assure myself it wasn't just me who was stressing out, everyone was stressed, and they're all shocked.

Student: I lost confidence in myself, and even forgot things that I already knew ... I actually started feeling numbness in my fingers ... the experience was very stressful I just want to forget this stressful experience.

The students reported that due to students experiencing physical distress and crying during the test some teachers employed unethical practices such as giving hints, correcting wrong answers, and taking answers from one student to another without the student's permission. They reported that the cheating was not by their EFL teachers, but by other invigilators of the rSECEE. They also reported that the invigilators were giving incorrect answers to students.

Would you believe me, if I said our classroom on the day of test was just like a debating room. They [the invigilators] were finding high achievers and asking for their answer sheets and they took their answers without their permission to cheat for others.

There were funny things that happened during the test, they were cheating students, and they gave students the wrong answer, isn't that ironic!

In one of the second focus group interviews one of the students, who was described by her classmates as a 'high achiever', complained about teachers' unethical practices and how their acts were jeopardising her test score. She reported being angry because cheating would unethically inflate her test score. She wanted the answer sheet to represent her own efforts not those of others. She further emphasized that the matter of teachers cheating students is multifaceted because students cannot object to such unethical behaviour. All the students stressed that the consequences of rejecting the cheating includes teachers and students negatively labelling them (e.g., Snob or Miss Clever). This meant that students have no option other than to accept these unethical educational behaviours.

The teachers were actually cheating us, and I was crying and asked her to stop giving me the answers. She said, “you’re strange, I am giving you answers and you’re refusing them”. I told her “I want to pass this test with my own efforts, I want to earn this diploma with my own efforts not with someone else’s efforts”. Any student who doesn't give answers during the test is disapproved by not only the students but also the teachers and labelled as the ‘Snob’ or ‘Miss Clever’.

In response to the above quote, the other focus group participants collectively said “And this why we accept cheating”.

All the student participants emphasised how teachers’ unethical testing practices of cheating could jeopardise their efforts over the course of the year, indicating students’ awareness of test score inflation and its impact on learning. The students also strongly highlighted the issue of labelling and how it was clearly used during the testing.

Student (Focus Group 2): It is disappointing to spend a whole year working so hard and then a teacher would come and just take your answer for other students. But what is even worse is to have teachers with limited English arguing with you about your own answers, they would say “your answer is incorrect because so and so who are the ‘*smartest*’ don’t have this answer” [emphasis added by the student with her fingers quoting the words *smartest*].

In response to the above quote the other students expressed negative attitudes towards the issue of labelling, and how their test results may be used as a “means of comparing” with others; how their performance in the rSECEE is used as a labelling indicator of whether they are ‘smarter’, or necessarily ‘dumber’ (Marchant, 2004, p. 2). They argued that labelling lowered their self-esteem.

Student (Focus Group 3): Labelling some students smarter, clever doesn’t make me feel jealous, but makes me feel that I am the thick one...this feeling is very discouraging and makes me feel why even bother.

### 7.5.2.2 Students' Attitudes Towards the rSECEE and Grade 12 EFL Teaching

Some students strongly highlighted how the rSECEE limited their opportunities for learning English, particularly, learning English for language use; hence, an example of negative washback. Some emphasized the intensity of the rSECEE affect has constrained learning English in Grade 12 classrooms to passing the rSECEE; in other words, teaching to the test. Consequently, as argued by the students, a high score on the rSECEE cannot be equated with the level of language proficiency.

Student (Focus Group 1): The typical English learning and teaching in our classroom is to teach English for the test only.

Student (Focus Group 2): Learning English is linked with passing the test, not with how well you can speak it or write in English.

Student (Focus Group 3): A test score is not indicator of this person can actually speak or use English. 99% doesn't essentially mean they like or can speak English.

They further argued that if the Grade 12 EFL classes were tailored towards enhancing and promoting language, teachers would change both the content and method. They highlighted that neglected skills such as speaking, listening and writing, could actually be taught and be taught differently.

Student (Focus Group 2): The true objective of studying English in Grade 12 is to pass the test, but if the objective was the opposite i.e., learning English for actual use, then the teachers focus would be different. She would focus on speaking, listening and writing, and would also change her teaching style.

Moreover, one of the students made an interesting reflection, with which her classmates agreed. She reflected on the relationship between teachers' BAK (Woods, 1996), classroom practices, and the high-stakes and the type of washback on the students; hence, the teacher factor and washback (Watanabe, 2004).

Student (Focus Group 3): The teacher's perspective about the subject and how she perceives the final tests has a great deal of an affect on her students.

However, they later explained that the teachers' complete focus on grammar knowledge and vocabulary is justified because the rSECEE focuses on these areas. In addition, the prevalent EFL language learning culture (see Section 2.4.2) emphasizes "teach to the test" and gives preference to the acquisition of the linguistic rather than the communicative knowledge of a language.

Student (Focus Group 1): When we are working through the book [it would be good if] the teacher would say don't focus on this; *it* being either writing, speaking or listening -- because it is not the test.

Student (Focus Group 2): Because of the way English was taught to us over the years. What I mean by that is, it was taught just as any other subject, but we don't have good background in English.

Student (Focus Group 3): If we were educated and taught from the early years of schooling to appreciate language use and not just the test we would have never accept teachers to teach English to the tests only. From the start to Grade 12 we had seven yeas of improper language teaching.

Finally, when asking students what advice they would you like to communicate to Libyan policy-makers responsible for secondary level education system, their recommendations included: provide teacher development programmes that educate teachers about how to teach language use rather than teach to the test; revise the curriculum to serve the needs of each specialisation; provide teaching resources and public facilities that help support and promote language learning; revise the test so that it covers all content domain; involve Grade 12 teachers

during the test development process; and make the objectives of both curriculum and test transparent to the public. The following quotes echo the students' recommendations:

Student (Focus Group 3): They need to provide teachers with workshops, teachers need to know how to teach language for communication purpose, I don't think they have received training. By doing this we would learn how to use in English when we go abroad.

Student (Focus Group 2): They need to get Grade 12 teachers to set the test. Okay they may be scared of teachers cheating the students, but why not get a group of Grade 12 teacher to write test items and then the language inspectors can select from them each year.

Student (Focus Group 2): I want to tell them, I don't want a textbook, I want language and how to use language. That's all I want. So, fix the test to cover all of the curriculum.

Student (Focus Group 1): I would say the curriculum, they need to look into revising it. The topics are boring, why not change it to fit the needs of each stream. Like have English texts related to chemistry or biology, not like what we have today. We have readings on sports.

Student (Focus Group 1): The current government isn't supporting English language teaching to help teachers teach real English, to teach English for communication. They need to provide teachers with resources. Plus, the government needs to have public facilities to support language learning. If these facilities were available and accessible it would help improve Libyan students' levels of English.

Student (Focus Group 3): They need to make the objectives of the curriculum and the test clear and everyone can access them from the Ministry's website.

The student recommendations are considered in the concluding chapter of the findings in the following section.

## 7.6. Discussion

This research is guided by two overarching research questions: *What is the relationship between the degree of alignment and the washback of the rSECEE? What are the implications of this relationship for key stakeholders?*

To answer these overarching questions, other questions were posed, including the second research question of the study, which investigated the degree of alignment between the rSECEE and Libya's EFL standards. This was addressed in Phase 1 by looking whether rSECEE meets the Webb's model (1997) comprehensive criteria. The Phase I research question falls into the macro level context of high-stakes tests and its relationship with the other two educational components (standards and the curriculum).

The results of the Phase I study guided me to investigate the third research question:

**What is the nature and scope (Cheng, 2004) of the washback (if any) of the rSECEE on the Libyan EFL Grade 12 teaching and learning?**

This research question, which is the focus of the Phase II study, looks at the washback effect of high-stakes tests at the micro level (i.e., within the classroom). The Phase II research question was addressed by looking at any evidence of washback: (1) on how the rSECEE influences teachers' accounts of teaching and assessment; (2) to what degree the rSECEE appeared to influence teachers' teaching practices; (3) and how the rSECEE influences learners' accounts of learning.

Having presented the study's findings in Section 7.5, this section discusses the findings in detail and in relation to the Phase II research question and research literature that framed this study. The gathered evidence falls into the following themes which are discussed in turn below.

1. The washback effect on teachers and teaching (content and instructional practices).
2. The washback effect on learners and learning.

Before commencing with the discussion, I would like to restate what was considered as evidence of washback in the current study. It was stated in Section 4.4, that washback in the current study permits for “both the accidental and intentional effects of washback and leaves the door open on whether washback is positive or negative” (Spratt, 2005, p.8). In addition, the study follows the approaches of Cheng (2005) and Messick (1996) for identifying washback, where Cheng (2005) contends that washback indicates “an intended or unintended (accidental) direction and function of curriculum change on aspects of teaching and learning by means of a change of public examinations” (p.112). Messick (1996) identifies the effects of tests as washback only when they can be linked to the introduction and use of the target test. Therefore, any effects, either positive or negative or either intended or unintended, linked to the introduction of the rSECEE in Libya are considered as washback in this study.

It is also important to acknowledge that making value judgements, either positive or negative on the impact of tests in general and the rSECEE in particular, is a complex and contested matter (Burrows, 2004; Cheng, 2004, 2005). In addition, identifying the evaluator is important whenever value judgements are passed about a test, because, as argued by Cheng and Curtis (2004) and Watanabe (2004), whether the washback is considered positive or negative depends on the evaluator’s perspective on what qualifies as positive or negative washback. Cheng and Curtis (2004) further argue that there is little agreement within the field of language testing on what test effects constitute positive or negative washback. In my researcher’s role, I consider myself to be outside the rSECEE testing context, therefore any evidence of washback is

evaluated either as positive or negative, in relation to Libya's EFL standards that are operationalised within curriculum and endorsed in the national high-stakes test and the Grade 12 EFL classroom.

### **7.6.1 The Washback Effect on Teachers and Teaching**

The reported results on the washback of rSECEE on teachers and teaching indicate that the implementation of the rSECEE has had mixed effects; it has had a negative washback on some teachers and their teaching but not others. From the findings it can be argued that in a high-stakes testing context, such as Libya, the operating test may affect some teachers and their teaching (content and practices), but not others, hence supporting Alderson and Wall's (1993) hypotheses, namely: "[t]ests will have washback effects for some teachers, but not for others" (p121). In addition, the study's findings echo Hughes' (1993) proposed washback mechanism and Green's (2007) washback model. According to Hughes' (1993) model, the nature of the test may directly influence the perceptions and attitudes of its participants towards teaching and learning. Accordingly, these perceptions and attitudes may then affect what the participants do (see Section 7.6.2 for further discussion). Similarly, Green's (2007) model considers participants' attitudes and values as one of the important factors that "will moderate the strength of any effect, [hence, intensity], and, perhaps, the evaluation of its direction" (p.25). In essence, teachers' views towards the rSECEE may in turn affect the direction and intensity of its washback, and how the resulting washback is evaluated.

#### **7.6.1.1 The Washback Effect on Teachers**

Negative washback on Teachers A and B included teacher dissatisfaction because the rSECEE constrained their opportunities to practice communicative activities and exerted considerable stress due to their accountability towards the students, parent, and the school. The

feeling of accountability and anxiety had consequential effects on Teachers A and B, with the former temporarily opting out from teaching Grade 12 students, and the later having health issues. It is important to note that the type of teacher-learner accountability portrayed by the teachers in this particular research context differs from teacher-system *accountability*. In other words, whereas, the former is a teacher's concern towards students' learning and development (Cook-Sather, 2010), the latter, is the type embedded in accountability systems where teachers and schools are rewarded or sanctioned for students' achievements measured by tests scores. (Lingard, Sellar, & Lewis, 2017).

Teachers have also reported that because of the stakes associated with rSECEE and the pressure of producing high test scores they have a “diminished sense of professional worth and feeling of disempowerment and alienation” (Blazer, 2011, p.6). Teachers A and B described themselves as the “spoon-feeders” of information, and stated that their professional roles were restricted to fully preparing students for the rSECEE and not for teaching English language or its use.

In contrast, the rSECEE did not appear to have negative washback on Teacher C. The findings indicate that the rSECEE caused no feeling of anxiety or stress for Teacher C. The teacher's sense of relaxation was reflected in her classroom learning atmosphere with it being full of laughter and excitement. Teacher C's stress-free attitude directed her sense of accountability towards encouraging higher order thinking skills, such as critical thinking, in her students rather than teaching to the rSECEE.

Furthermore, the findings indicate that all three teachers had negative sentiments in terms of the rSECEE's limitations. These included the rSECEE being an: inaccurate indicator of success; achievement test rather than a proficiency test; and “unfair” instrument for evaluating Grade 12

students. The inadequate time for students to practice and students' language levels were two major worries that the teachers repeatedly reported as problems with the new curriculum.

Thus, it can be argued that the high-stakes nature of the rSECEE meant that “success in the exam was a way to judge teachers' professional value” (Tsagari, 2011, p.438). This, in turn, made some Libyan teachers feel accountable to a number of stakeholders (such as students, parents, principals, and the school on the whole), and led to high levels of stress and anxiety. This mirrors the findings of Tsagari (2011) where teachers equated students' success on the First Certificate in English examination (FCE) with their individual professional value and success.

Moreover, following Anagnostopoulos (2003), Au (2007) and Lobascher (2011), it can be further argued that testing (such as the rSECEE) may undermine teacher creativity. Teachers become “technicians” (Hargreaves, 1994) where the fundamental love of learning in students is removed and replaced by extrinsic rewards and threats that, in turn, reduce enjoyment of the teaching and learning experience (Polesel et al., 2012). The following quotes from the teacher participants vividly illustrate this claim:

Teacher B: I am not happy about my current teaching. Instead of actually teaching a language, what I do is just instill information to students in order for them to pass their test.

Teacher C: The whole thing is wrong, students are no longer thinkers since the implementation of this form of testing. Our students no longer have the motivation to learn or further their knowledge in different areas of interest. We no longer have thinkers or analysers in our schools ... Because of our imbedded testing system, we no longer have open minded and reflective students.

A large volume of research, within both the educational and language testing literature, documents that the negative attitudes that tests can engender among teachers, which include feelings of frustration, high anxiety levels, and a decrease in faculty moral (Alderson & Hamp-Lyons; Blazer, 2011; Linn, 2006; Onaiba, 2014; Shohamy et al., 1996; Taylor et al., 2002).

Likewise, the current study found support for variability across individual teachers: some teachers develop negative sentiments; some do not. This difference possibly arises because of the teachers' differing beliefs systems and how these shapes how innovative information is acquired, acknowledged, and ultimately, responded to (Borg, 2001).

The following sections examine how teachers' attitudes towards the rSECEE "impact on *what* teachers teach" (Wall & Alderson, 1993, p. 63) and *how* they teach it.

#### **7.6.1.2. The Washback Effect on Teaching Content**

As reported in Section 7.5 the rSECEE had a mixed impact on teachers' choices of instructional content and material. The rSECEE had negative washback on Teachers A and B Grade 12 EFL classroom teaching content, which was mainly tailored towards the rSECEE's testing requirements. The delivered content in their classes resulted in the curriculum being narrowed to only grammar and vocabulary. Grammar content was delivered in isolation or integrated with content knowledge of the reading texts and vocabulary. Teachers narrowed the curriculum by ignoring the actual teaching of any content domain related to speaking, writing and listening because it did not contribute to the testing requisites of the rSECEE.

However, the negative washback of rSECEE on teachers' teaching content was variable, because Teacher C showed little-to-no evidence of narrowing the Grade 12 teaching content. Instead her classroom teaching content was guided by the textbook objectives which aimed at promoting communicative competence and knowledge of the language. Thus, the teaching content was operationalising the Grade 12 EFL textbook objectives rather than the testing requisites of the rSECEE.

Data drawn from the classroom observations contributed significantly to the findings on how the rSECEE affected the teaching content of Grade 12 EFL teachers. The findings indicated that despite an established curriculum (standardised textbook) that encourages critical thinking and creativity, and provides ample opportunities for implementing language use, some teachers narrowed the curriculum, although others did not. For Teachers A and B content was “an area of high washback intensity” (Cheng, 1997, p. 50), whereby they narrowed the focus of the pedagogical attention to the curriculum components (mainly grammar and vocabulary) that were to be tested, while other untested content (i.e., writing, speaking and listening activities) was marginalised. Thus, it was found that certain Libyan Grade 12 teachers were inclined to model their classroom learning activities to mirror the rSECEE’s form. One explanation for the rSECEE having a such powerful affect on content was explained by both Teacher A and B as the concern that stakeholders, such as the students and their parents, would complain if the classroom practices were not aligned with the rSECEE.

The narrowing of the teacher’s instructional content occurred in two ways, in both the content that was selected, and the way it was presented -- a process of *narrowing* similar to that reported by Au (2008). In other words, only the content on the rSECEE was viewed as “legitimate knowledge” (Polesel, 2012, p.11) – all other content (even that which was clearly defined by the curriculum/textbooks) was marginalized. In addition, teachers presented the content concurrent to the requirements of the rSECEE, by fragmenting and isolating knowledge forms into more discrete, test-driven forms. Importantly, the teachers’ narrowing of the curriculum did not necessarily mean that the teachers were insensitive to their students’ learning or did not invest effort in promoting learning. However, similar to Abobakeer (2017), the question that should be asked, is how well the students have mastered the target language

content, and whether the students would be able to apply the knowledge or skills in a *real-life* context.

It can be further argued that any narrowing of the curriculum and the domination of day-to-day activities that closely resemble the test format may deny students learning opportunities in many ways. First, learning isolated facts and skills can be difficult, because without context there is no meaningful way to accumulate or systematize information and make it easy to remember (Shepard, 1991). Second, learning “decontextualized skills means that the subsequent application of skills to real world problem becomes a separate and difficult learning hurdle” (Shepard, 1991, p.233). Lastly, learning itself may be devalued and discourage students, especially those who are most in need of improvement (West, 2007).

In summary, the results in part echo the substantial volume of high-stakes testing and washback research, including, but not limited to, Amrein and Berliner (2002), Cimbricz (2002), Erfani (2013), Freeman (2004), Hayes and Read (2004), Shepard (2002), and Wall and Alderson (1993). These researchers agree with the hypothesis that a “test will influence what teachers teach” (Alderson & Wall, 1993, p.120), as asserted most pointedly by Valli and Buese (2007): “curriculum coverage, matching taught content to tested content, and finding appropriate materials for students become [a teacher’s] overriding considerations” (p. 456).

However, the findings of the current study also provide evidence that challenges the above claims and supports an alternative view. In accordance with Gradwell (2006) and van Hover (2006), I argue that high-stakes testing, such as the rSECEE, may have little to no effect on what teachers do in the classroom, hence validating another one of Alderson and Wall’s (1993) washback hypotheses that “[t]ests will have washback effects for some teachers, but not for others” (p.121).

### 7.6.1.3 The Washback Effect on Instructional Practices

The findings from both the teachers' accounts and the classroom observations indicated that the effect of the rSECEE on instructional practices was far reaching for some teachers, but not for all. The rSECEE had a negative washback effect on Teachers A and B's Grade 12 EFL instructional practices, as it encouraged them to adopt approaches they would not otherwise have adopted. Teachers A and B devoted almost all of their teaching time to "teaching to the test"; instructional time was increased for and limited to the rSECEE tested skills (e.g., language rules and vocabulary) at the expense of the untested skills (such as speaking, listening, reading and writing).

The marginalised content was taught deductively<sup>36</sup> and in a decontextualized manner resembling the rSECEE format, hence employing the grammar translation method for teaching grammar and vocabulary. Teachers A and B favoured the grammar translation method "because it was the most appropriate method to cover the curriculum and meet the test demands within the time frame" (Teacher A). In other words, classroom activities (such as individual seat work, and teacher-centred grammar explanations) that are central to the grammar translation method were justified because they were compatible with the principles underlying the rSECEE. Grammar and reading were not taught to promote students' communicative competence.

In particular, Teacher B, due to her BAK (Woods, 1996) which to a degree oriented her towards communicative teaching approaches, paid minimal attention to speaking, listening, reading which were taught through pair- and group-work activities. Teacher B's accounts indicated that taking chances with the communicative approaches was risky and could jeopardized students' success with the rSECEE: "I won't let this [failing the test] happen to my

---

<sup>36</sup> A deductive approach (rule-driven) starts with the presentation of a rule and is followed by examples in which the rule is applied (Thornbury, 1999).

students and I will do what is in their best interest to help them pass the test. So, what I do is, I prepare them very well for the test”. Essentially, the rSECEE for Teacher B was “the ferocious master of the educational process, not the compliant servant [it] should be” (Madaus, 1988, p.85).

Because the rSECEE plays a fundamental role in the Libyan context and the decisions related to their rSECEE performance carry important consequences for students, Teachers A and B engaged in test preparation activities and material to help Grade 12 students achieve the highest test scores possible. Accordingly, they may have been “exam slaves” (Lam, 1994, p. 91), wherein there seems to be “no interest in the thought processes, only in the performance” (Baird et al., 2017, p.319).

Nonetheless, Teacher C, had little-to-no apparent washback affect on how she delivered her Grade 12 EFL content because she did not believe in teaching to the test. Hence a teacher’s attitudes towards the rSECEE can affect the what and how of the classroom (Hughes, 1993; Green, 2007, 2013). The Grade 12 standardised textbook was operationalised in keeping with its communicative intent and then her classroom teaching and learning. In the observed classes Teacher C had no teacher-centered activities that focussed on the teaching of grammatical content. In addition, Teacher C taught reading and grammar content inductively,<sup>37</sup> and by imbedding content within other language skills. Reading was not taught in reference to the rSECEE, rather, Teacher C considered this skill as a means to an end; something to be used in order to understand, apply, act – as an outcome of reading a text. Moreover, she implemented student presentations. Teacher C regularly exchanged information with her students in English and was mainly concerned about developing their thought processes, rather than focusing solely

---

<sup>37</sup> An inductive approach i.e. rule-discovery, starts with some examples from which a rule is then inferred (Thornbury, 1999).

on products in all her language classroom activities that involved student interaction. Finally, Teacher C, as reported by herself and her students, did not employ test preparation practices in her classroom. Instead, her instructional efforts were concentrated on language use. In essence, her teaching style was not a “transmission style” but more or less a more problem-solving or “interpretative” style of learning that involves collaboration between students and teachers (Taylor, Fraser, & Fisher, 1997).

The findings in relation to the rSECEE and its washback on Teacher A and B’s instructional practice, reflect the various criticisms cited within the research literature. These large-scale high-stakes testing criticisms include:

- The potential risk of reducing students’ “opportunity to learn” and opportunities to engage with cognitively demanding tasks (Madaus, 1988; Luna & Turner, 2001; Resnick et al., 2004; Smith & Rottenberg, 1991; Stecher, 2002; Valli & Buese, 2007);
- The inconsistency of meeting the set aims (Roach et al., 2005; Roach et al., 2008, Webb, 2002, 2005); and
- The negative impact of testing on teachers and teaching, and learners and learning (Brimijoin, 2005; Cheng, 1998, 2008; Green, 2013; Lee, 1994; Luna & Turner, 2001; Roach, Niebling, & Kurz, 2008; Smith, 1991b; Rottenberg & Smith, 1990; Smith & Rottenberg, 1991; Shohamy et al., 1996; Stecher, 2002).

In addition, to prepare students for the high stakes teachers, such as Teachers A and B, may use teaching methods that abandon or limit “innovative instructional strategies such as cooperative learning” in “favor of more traditional lecture and recitation” (Blazer, 2011, p.3). The teachers’ exclusion of non-tested topics within subject areas and acclimating their teaching style to the testing format, and implementing test preparation practice may lead to test score

inflation, which, in turn, makes it challenging for those who base decisions on the results, as well as “for those interpreting results to identify meaningful gains in student learning” (Looney, 2009, p.11).

What was identified in some teachers’ classes (i.e., Teacher A and B) is the important status that the rSECEE has in the Libyan education system. Similar to the Hong Kong context studied by Cheng (2004), the power of the rSECEE has in part driven teaching in some teachers’ classes in the “direction of coaching and drilling for what” is necessitated in the test (p.164). Examples of the rSECEE driving force was shown in some teachers’ classroom activities that mimicked the test’s activities. Putting it differently, similar to other high-stakes testing programmes the rSECEE has sent very strong signals to some teachers about what they ought to be teaching and what students should be leaning, and, therefore, some teachers tend to teach what is being tested (Herman, 2004; Koretz, Mitchell, Barron, & Keith, 1996; Stecher et al., 2000). Essentially, as noted by Wall and Alderson (1993) “tests can be powerful determiners, both positively and negatively, of what happens in classrooms” (p.41).

However, within the same high-stakes testing context there are other teachers, such as Teacher C, for whom the high-stakes test has little-to-no negative washback on what teachers do in the classroom. Instead, in line with Smith (1991b), Teacher C resisted the pressure to narrow the curriculum to fit the rSECEE prerequisites. This may be due in part to the dissonance between the ideologies of teaching and learning implicit in the rSECEE and Teacher C’s BAK (Woods, 1996) – what Teacher C considers *good learning practices* and her view of what students *need to learn*. Therefore, when a teacher’s beliefs (such as Teacher C) regarding appropriate teaching practices are not in line with the apparent emphasis of a test, and where these practices “are valued more highly (or at least better understood) than success on the test”

(Green, 2007, p.22), washback would seem less likely. Thus, washback intensity (Cheng, 1997; Watanabe, 1996, 2004) on Grade 12 EFL teachers' instructional practices will vary from one teacher to another.

In a similar vein, Teacher C may have had to make choices about best classroom practices for promoting language learning that can achieve two purposes: the first being sufficient learning of the EFL curriculum and the second being good rSECEE results. Teacher C may have also valued her BAK about appropriate language practices over success in the rSECEE, and therefore, there was no evidence of the rSECEE's negative washback on her Grade 12 classrooms. Thus, the effect of testing on teachers' instructional practices, as emphasized by Spratt (2005), can be also seen as highly individual, even though the washback literature generally finds that a "test [will] have impact on *what* teachers teach but not how they teach" (Wall & Alderson, 1993, p.68). Following Spratte (2005), Wall (1999), and other washback research findings (e.g., Onaiaba, 2014; Wang, 2009; Yu, 2010), the results of this study suggest that the similarity between a testing instrument's philosophical underpinnings and a teacher's BAK is a fundamental and influential factor in determining the direction and intensity of the washback.

Therefore, it can be concluded that the degree and type of washback occurs through "the agency of various intervening bodies and [is] shaped by them", and an "important and influential agent in this process is the teacher" (Spratte, 2005 p.26). Along similar lines, Shohamy (1992) emphasizes the central role of teachers in determining the potential for negative washback.

It can be further argued that teachers within the rSECEE testing context "face a set of pedagogic and ethical decisions about what and how best to teach and facilitate learning" (Spratte, 2005 p.26). Teacher A and, in particular Teacher B, believed that the rSECEE only sampled a limited part of the Grade 12 content domain, and fully understood that by narrowing

the curriculum to the rSECEE's requirements they were not serving their students' long-term interests. Despite such awareness both teachers had to make ethical decisions regarding their approaches to Grade 12 instructional practices. They had to decide between covering the whole grade 12 EFL textbook or *teaching to the test*. The Teachers feeling a sense of accountability towards students and helping their students to succeed in the test appear to have helped them decide what could be conceived as best practice. In turn, this echoes Popham's (1987) accurately worded criticism that:

Few educators would dispute the claim that these sorts of high-stakes tests markedly influence the nature of instructional programs. Whether they are concerned about their own self-esteem or their students' well-being, teachers clearly want students to perform well on such tests. Accordingly, teachers tend to focus a significant portion of their instructional activities on the knowledge and skills assessed by such tests. (p.680)

Although the Phase II results are complex, it can be argued that the existence of the rSECEE may have in part caused negative washback in many aspects of some Libyan teachers' learning and teaching practices. In essence, it was "testing, not the "official" stated curriculum, that [was] increasingly determining what is taught, how it is taught, what is learned, and how it is learned" (Madaus, 1988, p.83). However, it can also be argued that the presence of the rSECEE may have limited-to-no washback affect on what teachers want to do in their EFL classrooms. Instead, some Libyan teachers take full advantage of the Grade 12 EFL curriculum to improve Libyan students' proficiency in English for the purpose of authentic language use and have been inclined to resist the power and influence of the rSECEE (Burrows. 2004)

#### 7.6.1.4. Burrows' (2004) Framework of Analysis

Employing Burrow's (2004) analysis and interpretation of the observational data, the findings are next interpreted through the lens of curriculum innovation within the field of educational change. Lambright and Flynn's (1980) distinctive stakeholder roles are used to analyse the roles of the three Libyan EFL teachers in relation to the rSECEE. Following Burrows' (2004), I placed the three teacher participants on a continuum (Table 7.5) from a teacher who felt deeply affected by the rSECEE (Teacher A, the *adopter*) to one who felt least affected (Teacher C, the *partial adapter*). Teacher B (the *adopter*) was in between the two, but much closer to Teacher A than C, as indicated in Table 7.5 below:

These findings indicate that the washback of rSECEE is a highly complex matter. In accordance with current washback research, for example, Burrows (2004), Wantanabe (2004) and Spratt (2005), how teachers react to rSECEE is individual. Similar to Burrow's (2004) results, it can be argued that there is an implicit continuum from Teacher A, who appears to have experienced considerable, observable negative washback from the rSECEE on her Grade 12 teaching, to Teacher C who appears to have little-to-no observable effect from the rSECEE on her teaching. In between the two teachers, we have Teacher B, who appears to have experienced observable negative washback from rSECEE on her teaching, but not to the extent that of Teacher A. Consequently, this leads to a categorisation that is parallel to Burrow's where Teacher A and B are labelled as adopters, although Teacher B may be seen as a partial adopter. Teacher C, however, may be seen as a partial adapter. An adapter is one who takes into the account the idea of "choice", i.e., as a participant of a social system who takes from "the new system as she or he chooses" (Burrows, 2004, p.125), and it is a teacher's BAK (Woods, 1996) that "serves as a guide to thought and behaviour" (Borg, 2001, p.186).



esteem. In addition, negative washback on students was also reflected in students' attitudes. The students used similar language to their teachers when expressing frustration with the rSECEE and its reliability and validity. Grade 12 students' negative sentiments towards the rSECEE included it having limited coverage of the set domain of knowledge, which in turn lead some Grade 12 teachers to narrow the curriculum. Teacher C's students did not complain of any narrowing of the curriculum in their classroom. Their complaints, however, were towards the rSECEE not measuring the language abilities that were practiced in class; having misleading test items that encouraged guessing; and providing an inaccurate measurement of their true language ability. Finally, the rSECEE was described by the students as being an *unfair* form of testing because both guessing and cheating were inflating their test scores and determined success.

Student (Focus Group 3): I can't see it as a fair test, because there is cheating and guessing. Some students can pass the test just by guessing some questions and the teachers cheating them.

Therefore, these findings mirror what is regularly reported in the literature (e.g., Fox & Cheng, 2007; Hawkey 2006; Shih, 2006) that students' fear of high-stakes testing and its consequences is probably due to what is at stake (Paris, 2000). As the rSECEE's stakes become more "consequential for students" as is the case of Libyan students, the pressure to do well may increase correspondingly (Paris, 2000, p.3). In this context, Hawkey (2006) emphasizes that it is important to identify if students' fears play a role in underperforming on tests. If this is the case, then these factors may provide inaccurate results about a student's actual ability.

The study's findings provide further support to the significant role teachers play in mediating washback effects on students (Hamp-Lyons, 1996; Cheng, 1999, Shohamy et al., 1996; Spratte, 2005; Watanabe, 2004; Yu, 2010). In addition, the findings related to Libyan students' experiencing emotional, psychological and physical distress reflect the far-reaching impact of

high-stakes testing for students in that it can negatively affect students' health and well-being (Polesel et al., 2012). In turn, this confirms international research findings conducted in, but not limited to the US, UK, China, Singapore and Nepal, that high-stakes testing can affect students' mental and physical conditions (see Section 4.4). In particular, the findings reflect Marchant's (2004) accurately worded criticism that by the time the high-stakes results are published, some students may use them as "means of comparing" themselves with others; hence, the test is used as a labelling indicator of whether they are "smarter", or necessarily "dumber" (p.2). Moreover, the findings about the doubted fairness of rSECEE and its results suggest that "the way in which learners respond to an assessment has a good deal to do with the design of the assessment and the content it covers" (Green, 2014, p.87), and that washback on learners is mediated through teachers (Scott, 2005; Yu, 2010).

#### **7.6.2.2 The Washback Effect on Student Learning**

The findings reported in Section 7.5 indicate that the rSECEE had a negative washback on both teachers' and students' attitudes towards student learning. All three teachers stressed that the revised testing system had marginalised language learning. Teachers A and B emphasized that the only type of student learning being promoted and catered for in their individual Grade 12 classrooms was memorisation -- the skill that was most useful for performance on the rSECEE. Teacher B further criticized the rSECEE for undermining student autonomy, limiting language use and impeding the promotion of students' communicative competence in Grade 12 classroom.

Teacher C emphasized a similar perspective to her colleagues. Her accounts of the shortcomings of the revised testing system were not a result of her reflecting on her current Grade 12 classroom, but rather a reflection on the whole secondary level education system. In

contrast to other teachers, Teacher C felt that actual language learning was taking place in her Grade 12 classrooms. By building a good teacher-student rapport and through the constant encouragement of students' abilities, her students were able to experiment with English and practice within the classroom learning environment.

It can be further concluded from the classroom observations of Teacher A and B and their students that the enhancement and development of students' linguistic knowledge was not effective for language use. On a number of occasions students unsuccessfully experimented with the acquired grammatical knowledge and tried to communicate knowledge using those forms. Actual language use, however, turned out to be a "separate and difficult learning hurdle" for them (Shepard, 1991, p. 233). However, this was not the case with Teacher C's students, as they were able to put the acquired language knowledge to use during group presentation and when communicating in class (e.g., requesting information). Importantly, the rSECEE multiple-choice responses, although they may provide some information about Libyan students' Grade 12 content knowledge, as a one-time paper and pencil assessment, the test has "serious limitations in measuring the variety and scope of classroom learning" that may have occurred in Teacher C's Grade 12 classrooms (Marchant, 2004, p.3).

Interestingly, all the students shared similar views to their teachers that the revised testing system had marginalised learning and limited learning to lower-level thinking skills such as memorisation and the recall of information. Although Teacher C's students did not appear to experience narrowing of the curriculum, or teaching to the testing requirements of the rSECEE, they still believed that the rSECEE had marginalised their learning by forcing them to memorise and recall isolated discrete information and knowledge. As one student noted, "the true objective of studying English in Grade 12 is to pass the test" and that "we just memorise and memorise

and that's it and nothing more". Madaus (1989) explains: "If students, teacher, or administrators believe that the results of an examination are important, it matters very little whether this is really true or false—the effect is produced by what individuals perceive to be the case" (p.88). This may be an important factor in understanding the washback of the rSECEE on Grade 12 EFL teaching and learning, which adds support to the notion that the:

[l]ack of 'fit' between the 'users' (learners) and the assumptions of the innovative methodology was [a] result of 'value conflict', [and] learners' beliefs and assumptions about the norms of appropriate classroom behaviours, which were entrenched in the culture of the community, clashed with the assumptions of the innovative methodology (Shamim, 1996, p.119).

Thus, the findings indicate that the former testing systems' shortcomings, which were highlighted by Al-Buseifl (2003), Alhamli (2007), Onaiba (2006, 2014), Orafi and Borg (2009), are all evident in the revised testing system such as the rSECEE. The shortcomings of the former testing systems, as discussed in Chapter Two (see Section 2.4), included the emphasis on assessing rote memorisation of vocabulary and grammar rules, and students' ability to recall of information, rather than on their ability to integrate and produce meaningful English in communicative situations. Similar to the revised BECE of English examined by Onaiba (2014), the rSECEE continued to encourage students to memorise extensive amounts of information and to perceive education as a process of only conveying information, rather than to develop their own explanations, to reason, or to draw conclusions (Abdulhamid, 2011). It can be further argued that the rSECEE has failed to achieve the Ministry of Education's goal of reducing the possibility for Libyan students to attain marks by fraudulent means, and, therefore, may have failed at augmenting the face validity of the revised testing system.

Moreover, based on the effect of the rSECEE on learning and supported by the research cited in the education literature and language testing literature (such as Cheng, 1998, Marchant 2004; Paris, 2000; Smith 1991a), it can be argued that the rSECEE is probably doing very little in improving students' knowledge and skills in English. In addition, the findings support Marchant's (2004) perception that the testing process of the current high-stakes test is an activity in which students demonstrate their knowledge and skills rather than their learning.

The above arguments add support to Hughes' (1993) washback trichotomy, namely, that a test will first influence the participants' perceptions, which in turn will influence their practices, and then the process of learning. The findings also mirror Resnik and Resnick's (1990) carefully worded critique of high-stakes testing programmes:

Standardised tests fare badly when judged against the criterion of assessing and promoting a thinking curriculum. They embody a definition of knowledge and skills as a collection of bits of information, and they demand fast nonreflective replies. The test and the classroom practices that might be used to prepare for them suggest to students a view of knowledge counter to what the thinking curriculum seeks to cultivate. (p. 73)

In addition, the poor item quality of the rSECEE (e.g., ambiguously worded test items, questions without a stem, tricky items, items testing too much information and not reflecting a specific content, and spelling errors), as noted by the three teachers and their students, may incorrectly inflate Grade 12 Libyan test scores for students who actually may not know the information tested by the question. The scores may be inflated because poorly crafted items can encourage students to guess (Downing, 2002), and thus contribute to CIV. Consequently, it may

be test-wiseness,<sup>38</sup> rather than learning that determines some Libyan 12 Grade students' success. On the other hand, students with ability to communicate in English may find their scores deflated because there is no clear answer to an item and they too are forced to guess or leave items blank.

Because of the importance of high stake tests, some Libyan students may have come to undervalue both schooling and learning, and for some the mere focus of schooling may have become “whether this will be tested or not” (Paris, 2000). There was no space or time given in Teacher A and B’s classroom for students to work in groups and discuss meaning. In essence, they were not given a “chance to broaden their perspectives and sharpen their understandings as they compare their ideas with others and make meaning” of classroom material (Abubaker, 2017, p.18). Since students have learnt to focus only on the tested information, it may be argued that, similar to other high-stakes tests reported by Paris (2000), the rSECEE has come to define the focus and learning of the Libyan secondary education system. Importantly, when students believe that a test defines the focus of learning and test scores are so important for future prospects, differences between high and low achievers in terms of motivation, anxiety and perceived self-competence may be great (Wigfield & Eccles, 1989).

### **7.7 Merging the Phase I and Phase II Findings**

In order to answer the main research questions —**What is the relationship between the degree of alignment and the washback of the rSECEE? What are the implications of this relationship for key stakeholders (e.g. policy- makers, test developers, teachers, and**

---

<sup>38</sup> Bachman (1990) includes test-wiseness as a personal attribute, i.e. characteristics that a candidate develops to assist them in writing a test including aspects such as guessing strategies and test pacing (p.114).

**students)?**— the following discussion synthesizes and merges the Phase I and Phase II results (Creswell, 2000).

Given that “it is not possible to form a direct one-to-one relationship between positive or negative washback and the perceived quality of a test” (Booth, 2012, p.41), and in accordance with Wall (2000, 2012), it would be an over-generalisation to base value judgements only on test design, as “test design is only one of the components in a quite complicated equation” (Wall, 2000, p.502), accordingly, the rSECEE’s negative washback on some Grade 12 EFL teaching and learning cannot be solely attributed to the direct influence of the rSECEE and its lack of alignment with the set EFL standards. Although, the rSECEE did have an affect on both Grade 12 teaching and learning practices, other factors operating within the Libyan educational context may also play an influential role in bringing about negative washback in some Libyan Grade 12 classrooms.

For example, the focus on linguistic knowledge rather than communication in teacher-centred classrooms, with teachers being suppliers of information, ‘teaching to the test’, and emphasizing rote learning and memorisation, may also be the result of the prevailing Libyan *culture of learning* (Cortazzi & Jin, 1997). In addition, this study suggests the influential role teachers play as change agents within an education system and how their BAKs in relation to the rSECEE may have played a role in determining the test’s washback direction and its intensity. Thus, the washback of the rSECEE is not just a simple cause and effect relationship between test and classroom experience, instead it is a broad, multi-faceted and “complex and contentious” (Booth, 2012, p. 41) phenomenon that can vary in both form and intensity (Cheng, 2005). Besides, it is a “phenomenon that does not exist automatically in its own right but is rather one that can be

brought into existence through the agency of teachers, students or others involved in the test-taking process” (Spratt, 2005, p.21).

The following section discusses how (1) the rSECEE’s test design and its lack of alignment with Libya’s EFL content standards, (2) the prevailing culture of learning, and (3) the role of the teacher, all play a role in bringing about negative washback of rSECEE on *some* Grade 12 EFL teaching and learning.

### **7.7.1 The Lack of Alignment of the rSECEE and Washback**

Without a semblance of alignment, nothing will cohere in an education system (Biggs & Tang, 2001; Biggs, 2003; Näsström, 2008), with possible negative consequences for the whole system including negative washback. Similarly, within the washback research literature it is argued that the test format and the tested skills are factors in an education system that contribute to the complexity and unpredictability of washback (Green 2007, 2014; Shohamy et al., 1996; Wall, 2012). Furthermore, a well-designed instructional programme considers the relationship between the focal construct and the content of assessment which, in turn, as argued by Green (2007, 2014), may have an effect on the type washback that is experienced at the classroom level. Given this, the degree of alignment between the rSECEE’s test design and the skills developed by the Libyan EFL standard may have an effect on the type of washback operationalised in some Grade 12 EFL classroom teaching and learning. Therefore, the Phase I study looked at the degree of alignment between the rSECEE and Libya’s EFL content standards. Chapter Six discussed how the rSECEE has a limited degree of alignment with Libya’s EFL content standards, which, as argued by the current study, may have played a part in bringing about the negative washback that exists in *some* Libyan Grade 12 EFL classrooms. Each of the shortcomings that contributed to the rSECEE’s lack of alignment, and how these

may have encouraged the direction of negative washback on some Libyan Grade 12 classrooms is explained below. The shortcomings in terms of the rSECEE are: (1) emphasizing certain standards while ignoring others; (2) demanding only DOK Level 1 from students; (3) covering a limited range of the Grade 12 EFL content domain; and (4) having limited-to-no balance of representation of the content domain.

The rSECEEs give different weightings to the Libyan EFL content standards by varying the number of items measuring content related to certain standards (mainly Standard 1 and Standard 2). Thus, the Libyan EFL content standards have not been adequately operationalised within the rSECEE, thereby creating a very low level of agreement on Webb's (1997) categorical concurrence category. Accordingly, and as is evident from the observational data, this shortcoming may have encouraged Libyan EFL teachers, such as Teachers A and B, to focus their instructional practices on encouraging learning opportunities for Grade 12 students to attain Libyan EFL Standards 1 and 2 that reinforce and emphasize the importance of promoting linguistic knowledge. More specifically *some* Libyan teachers may have solely focussed on assisting students to “manipulate the English lab [the students’ linguistic knowledge of the language] as a linguistic system and to have some conscious knowledge of how it works at the level of phonology, morphology, syntax and discourse” (Libyan EFL Content Standard 1, 2016, Personal Communication with Ministry Officials). In addition, *some* Libyan teachers may have then also focussed on providing Grade 12 students’ lexical systems [students’ knowledge of vocabulary] with “words so that the students can discuss topics related to their specialization” (Libyan EFL content Standard 2, 2016, Personal Communication with Ministry officials).

An additional limitation of the rSECEE can be highlighted in terms of its 60 test items that only necessitated low-levels of cognitive demand (i.e., DOK Level 1). All these test items

demanded simple recall of information or recognition of facts, hence achieving a low DOK consistency on Webb's alignment model. Therefore, I argue that the rSECEE's test items that are only recall-type questions merely emphasize the importance of rote memorisation of textbook content. The students' memorisation of content is of doubtful importance to them and their teachers, and may further encourage the prominence of memorisation within Libyan educational culture. Therefore, the test design inadequacy may have encouraged some Libyan teachers (such as Teachers A and B) to focus on the importance of memorisation and the enhancement of DOK Level 1, rather than DOK Level 2 (skill concept) and Level 3 (strategic thinking) as operationalised by the majority of the Grade 12 EFL content objectives. The focus on DOK Level 1 in some Grade 12 EFL Libyan classrooms may have denied students from opportunities where meaningful aspects of thinking and learning are put into practice.

Moreover, the rSECEE did not cover the full range of Grade 12 EFL content objectives stipulated by the Libyan EFL standards; in other words, it tested a limited content of the Grade 12 EFL curriculum, hence not meeting Webb's ROK correspondence. By ignoring the curricular content related to speaking, listening, writing, the rSECEE limited its domain of evaluation to the curricular content of grammar and vocabulary. In addition, to the rSECEE having a weak ROK correspondence, its 60 test items were not evenly distributed across the objectives. Thus, there was a lack of balance of representation where the construct was grossly underrepresented because of assessing a limited portion of the target construct.

Accordingly, as frequently noted within the research literature and as was evident with the two classes of Teacher A and B, these two test design limitations, weak ROK correspondence and lack of balance of representation, may have encouraged some teachers to use instructional practices and materials that mirror both rSECEE's testing format and content (Dillon, 2006;

Jerald, 2006; Tienken, & Wilson, 2001). Thus, the rSECEE is covering grammar, vocabulary, and reading at the expense of other skills such as speaking, listening and writing, which were then untaught in *some* Grade 12 classrooms. The rSECEE breaks the measured language skills (grammar and vocabulary) into small, isolated, discrete-point and decontextualized elements rather than measuring all language skills in an integrated and cohesive manner. With rSECEE consisting of isolated decontextualized items it may lack authenticity because skills are assumed to be stable regardless of time and reason (Resnick & Resnick, 1989). As discussed in Section 7.6, the skills represented in the rSECEE have been evident and operationalised in *some* Grade 12 EFL classrooms. However, these instructional practices are not what the EFL standards had endorsed or the Libyan Ministry of Education had envisioned.

Lack of alignment between test content and the focal construct or target domain of knowledge may not only cause short-term washback (i.e., immediate washback on students' teaching and learning), but may also play a role in long-term negative washback on students. For example, if one standard is over weighted in the test, then the outcome may not accurately measure the target construct and grossly underrepresent the construct; thus, a case of CUR may be evident (Messick, 1989). Consequently, because some Libyan teachers teach to the test "incorrect inferences may be drawn about student achievement in the domain, because the test item sample does not adequately represent the population" (Downing, 2002, p.239). For example, university administrators at different natural science departments may interpret Libyan students' rSECEE test scores, as "indicating ability in all areas of the focal construct, [but] the scores may, in fact, reflect a relatively limited knowledge or ability" (Green, 2007, p.4).

If test scores do not accurately support the valid inferences about the subject matter being assessed, it may also result in test-score inflation which contributes to CIV and, therefore, "the

accurate interpretation of scores may be jeopardized”. In these circumstances, “the score on the test does not represent a random sample of the content domain to which one can draw legitimate inferences from test scores, rather its interpretation is muddled by interjecting non-randomness into the sampling” (Downing, 2002, p.239).

CIV, as noted by Jacob and Levitt (2003), can also be brought about by coaching and cheating on the part of the teacher. For example, the rSECEE teachers who were administering the test gave hints, corrected wrong answers, took answers from one student to another without the student’s permission, and assisted students during the tests. Thus, the outcomes of the rSECEE may not be compared to the level of achievement that the rSECEE is meant to represent. Furthermore, in an unaligned system, such as the Libyan secondary level education system, students, parents, educators, administrators and policymakers may be misled by reports and score inferences which mistakenly categorise students as high-achievers not because of their actual competence but rather because of CIV. The rSECEE results may not signify that students have attained the articulated standards and the outcomes may not respond to the deliberate actions of students and educators (Baker, 2005; Herman, 2004; Webb, Herman, & Webb, 2007).

With Libyan students not attaining the full scope of standards set by the Libyan Ministry of Education, they may arrive on local or foreign university campuses with “insufficient writing and oral communications skills to participate fully in academic programs” (Jamieson, Jones, Kirsch, Mosenthal, & Taylor, 2000, p.3). Unfortunately, this has been the case, from my personal experience as a graduate student and my observations working as a language teacher at a Canadian university. I can safely argue that many undergraduate and graduate Libyan students arrive in Canada with a language proficiency level that is not sufficient for academic success. To achieve an adequate proficiency level, the majority of Libyan students enrolled in LSASP in

North America and Canada had to take 12 to 18 months of intensive English language instruction. Therefore, “[d]eficiencies in alignment result in ambiguity that may affect some or all parts of the system, like an incubating virus dangerous but not obvious” (Baker, 2005, p.319).

Although this study is limited by its context, the analysis on how deficiencies in alignment may result in long-term negative washback on an education system provides support to Fox’s (2018, personal communication) notion of washback *magnitude*. Encompassing both washback variability and intensity, Fox’s (2018) notion of washback magnitude highlights the effect of testing on learners and learning overtime. Comparing the notion of washback magnitude to the fluctuation of earthquakes, Fox (2018) argues that the washback of testing on learners and their learning may be illusive when observed at a certain point in time. However, remarkable fluctuations of test washback may occur over time. For example, the washback of a university entrance language proficiency test would be of a lower magnitude if a student passed the test and entered university. However, the washback of the same test would be of a higher magnitude if a student was struggling with language, hence was unsuccessful at an English-speaking university programme.

In the context of this study, the washback *magnitude* of rSECEE may be seen as small because there was no difference between the numbers of Grade 12 students of Teacher A and Teacher C Grade 12 passing the rSECEE. In the long run, however, when students enter their university programmes the fluctuation of the washback effect is liable to vary. The long-term washback magnitude for students of Teachers A and B who taught to the test, may be higher than of the students of Teacher C, who promoted language use rather than teaching to the test. In contrast to students who were given a better chance for learning English and language use, Libyan students who were taught English according to the rSECEE requirements may find their

future real-world application of English language skills a “separate and difficult learning hurdle” (Shepard, 1991, p.233).

In a nutshell, it can be argued that in a context such as Libya where there is a limited-to-no degree of alignment between a high-stakes test and the content standards there is a possibility of negative washback occurring at the classroom level. Following Näsström’s (2008) analogy which compared educational alignment to a chain linking the three jewels (components) of an education system (see Figure 7.5), it can be argued that there is weak link in the chain within the Libyan secondary (see Figure 7.6).

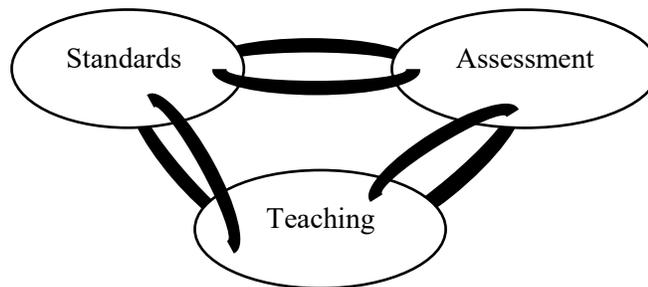


Figure 7.5: Alignment as Links between the Components of an Education System  
Source: Näsström (2008, p.20)

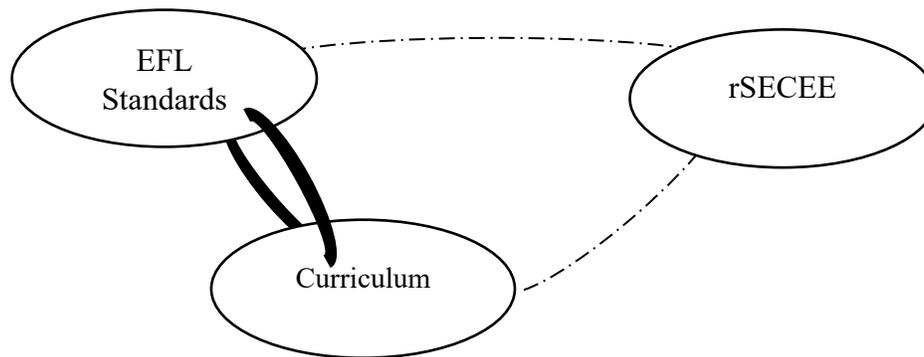


Figure 7.6: Alignment as Links between the Components of the Libyan Secondary Level Education System  
Source: Adapted from Näsström (2008, p.20)

In addition, the components (standards, curriculum and testing) within the Libyan education system may have drifted apart and may even possibly be giving totally different messages to both teachers and students about what they are expected to know and be able to do.

Moreover, as noted by (Biggs, 2003), in an unaligned system unintended messages may be sent to Libyan students who may understand that:

- Trees are more important than wood, i.e., success in the rSECEE is more important than learning English;
- Memorisation of isolated content gains marks; and
- Uncontrollable factors such as guessing and luck can contribute to success.

As a result of the weak alignment, Libya's educational components may be emphasizing different knowledge and skills, contradicting each other, be isolated from each other and in the long run may eventually break. Consequently, as uncovered by the current study, this may result in the secondary level education system operating less effectively, potentially increasing the level of anxiety and pressure for both teachers and students, and ultimately having a negative effect on Libyan society as a whole (e.g., if a student without an ability is admitted to university by virtue of a test score, and a student with the ability is not, particularly in settings with limited resources, this is a grave error for both the individuals and society at large).

### **7.7.2 The Libyan Culture of Learning Factor**

In Chapter Two the prevailing Libyan culture of learning in relation ELT was summarised as:

- (1) Emphasizes preparing students for high-stakes tests (i.e., is test—driven) (Al-Buseifl, 2003);
- (2) Places emphasis on memorisation of facts in classroom practices and testing (Alhmali, 2007);

- (3) Relies on grammar translation methods, e.g., of translation and the use of L1 (Altaieb 2013; Orafi & Borg, 2009)
- (4) Encourages teacher-centered activities, where teachers ask questions and students answer them (Altaieb, 2013; Omar, 2014; Orafi, 2008; Orafi & Borg, 2009) and leads to passive learners (Aldabbus, 2008);
- (5) Promotes students' linguistic competence (knowledge of language) rather than their communicative competence (ability to use language) (Omar, 2014; Onaibia, 2014; Orafi & Borg, 2009);
- (6) Correlates language learning with success in high-stakes tests (Abubaker, 2017);

It was further noted in Chapter Two that many Libyan researchers, including Omar (2014), argue that the revised secondary level curriculum may have not been adequately operationalised within secondary level EFL classrooms, because it had not shifted the classroom dynamics from teacher-centered to learner-centered pedagogical approaches. Fundamentally, as argued by Altaieb (2013) and Omar (2014), there is a mismatch between the objectives of the revised curriculum and the realities of the ELT classroom, which is also affected by a number of contextual factors, such as a lack of teacher professional development programmes, a lack of technical facilities and a lack of collaboration between schools and parents, teachers' beliefs, students' needs, school space, and time.

Considering the evidence for negative washback from the perspective of the prevailing Libyan culture of learning, one could argue that certain teaching and learning practices within the Grade 12 EFL classroom do not necessarily constitute washback, and that "the existence of the test [rSECEE] by itself does not guarantee washback" (Alderson & Hamp-Lyons, 1996, p. 281). Instead, and in accordance with Abdulhamid (2011) and Cortazzi and Jin (1997), the fixed

and embedded Libyan ‘culture of learning’ is a strong influential factor with regards to what happens inside a language classroom. In particular, the culture of learning determines the type of pedagogical approach to be implemented and the definition of what is considered to be a successful language learning environment. It also influences the beliefs, which are deep-rooted and entrenched within the whole educational system, about how teachers ought to teach and how students ought to learn. These beliefs are augmented over time; they begin on the day Libyans first walk into the classroom as young innocent learners, and continue to develop until the moment they graduate. Therefore, beliefs are not just rooted in teachers, but in the whole educational context, from policy-makers to students and their parents, and these educational norms have existed for decades within the Libyan context.

Moreover, it can be further argued that having memorisation embedded as part of the Libyan culture of learning may shape how the revised secondary level curriculum is delivered, because the ‘culture of learning’ is part of the “hidden curriculum” (Cortazzi & Jin, 1997, p.169) that each individual teacher uniquely operationalises in her/his classroom. Derived from the prevailing Libyan culture of learning this study adds further support to shortcomings of the Libyan ELT as highlighted above.

Consequently, it can be safely argued there is dissonance between the language teaching ideology embedded within the revised curriculum and the teaching and learning in the Grade 12 EFL classroom. Besides, “[t]esting cannot be neutral on what is taught and learned. Any test is an expression of values on teaching and learning” (Cole, 1999, p.1). Therefore, what is mostly valued by the Libyan culture of learning, memorisation and the enhancement of students’ linguistic knowledge, has been operationalised within the rSECEE, and together with the

rSECEE's test design limitations, may be considered a powerful contributor to negative washback.

### 7.7.3 The Teacher Factor

According to the educational literature, teachers' belief systems are a critical consideration in the introduction of any form of educational *change* (Hargreaves, 1989, 1993; Markee, 1997; Woods, 1996). Similarly, Burrows (2004), Green (2007) and Watanabe (2004) argue that teachers' beliefs about a change in an existing test, or the introduction of a new test, and what is considered best classroom practice, all play a role in both the direction and intensity of washback at the classroom level. Importantly, these notions move washback from "a recipe for achieving positive washback, towards a descriptive and partially explanatory tool addressing what goes on in order to cause the various washback effects" (Saville, 2009, p.32). It is not surprising that the washback effect of the revised testing programme is "not clear cut" but rather has "varying effects depending on the teacher involved" (Burrows, 2004, p.119).

Furthermore, following Minarechová (2012) it can be argued that Libyan language teachers and their students may have found a way around testing programmes through a hidden curriculum. By means of the hidden curriculum, the teacher participants may have made hypotheses, tested them, and then formed their own personal constructs. The constructs are their theories and beliefs, which may change and adapt with experience. Importantly, teachers' roles within an unaligned educational context such as Libya's may be seen as more influential, because they pull the educational components together according to their beliefs system and work around the misaligned system, tailoring it in what they deem to be the best interests of their students, to assure their students' success. Figure 7.7 illustrates how the weak links between the

education components presented above in Figure 7.6 are brought together through teacher mediation, in accordance with teachers' beliefs systems.

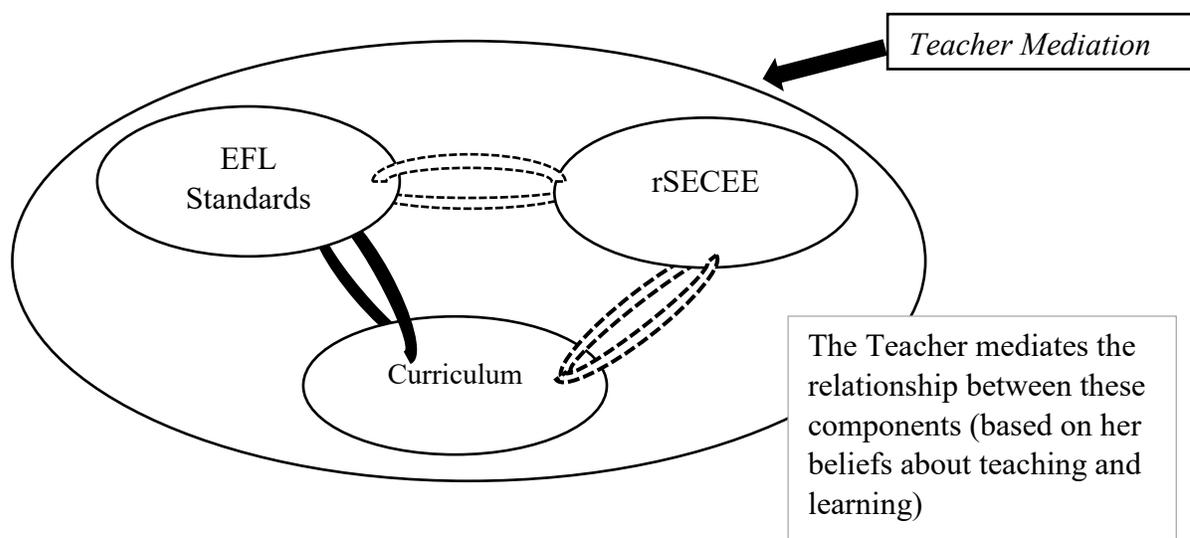


Figure 7.7: Alignment as Links Formed by Libyan Teachers between the Components of the Libyan Secondary Level Education System

Source: Adapted from Näsström (2008, p.20)

Importantly, the hidden curriculum does not “exist *a priori*. It is forged in the day-to-day interactions between students, their teachers, and their peers” (Booher-Jennings, 2008, p.150). Operating at all times, the hidden curriculum in Teachers A and B’s classes may have served “to transmit tacit messages to students about values, attitudes and principles” of the rSECEE and Grade 12 EFL teaching (Kentli, 2009, p.88). In essence, through the hidden curriculum, the values in relation to language learning were communicated between Teachers A and B and their students, reflecting the Libyan ELT *culture of learning*, wherein the teacher’s role is to “make the ... [student] understand” rather than to promote learning (Wall, 2007, p. 147).

In addition, through the hidden curriculum, the attitudes towards the rSECEE and how it drives both teaching and learning within the classroom is also communicated between the teachers and their students. Because the “[r]elationship between practices and principles is likely to be interactive; each will influence the other as the teacher works from day-to-day” (Breen,

Hird, Milton, Oliver, & Thwaite, 2001, p.472). Sources of the hidden curriculum defined in Teacher A and B's learning activities include the rSECEE, past test papers, the rSECEE content that emphasizes linguistic knowledge, language skills and methods that contribute to the requirements of the rSECEE.

In contrast, language teachers like Teacher C and her students may have learned to find their way around the rSECEE and resist its power over their EFL classroom by virtue of a different hidden curriculum, that favoured conditions of cooperation, trust, reflection, and thoughtfulness. Values were fostered in relation to language teaching which included the importance of promoting communicative competence and creating opportunities for language use of all four language skills (which were given equal weight in the classroom). The following quote from Teacher C, which she emphasized because she shifted to English (rather than Arabic) illustrates this:

Teacher C: I just want to make English as fun as I can, and I want them [students] to use English in their lives and I try to make it very interesting. I always tell them it is as if *we* are playing a game...I say to my students learning English is yummy, as yummy as a chocolate cake. (Emphasis added)

Values with respect to learning include shifting the conception of learning beyond rote memorisation of facts and measures to learning as an adventurous, engaging experience that shifts the passive learner role to that of active enquirer. The pronoun 'we' (rather than I) in the above quote suggests that both teacher and learner play important roles in the learning context.

To conclude, although this study has its limitations (see Chapter Eight) and many factors come into play in assessing washback, the results provide support for the notion that there is a "degree of choice involved in washback" (Burrows, 2004, p.125). Because "[i]f it is possible to choose to resist the effect of an implementation upon one's teaching, then it is possible to choose

whether the implementation of an assessment system or test will have a washback effect” (Burrows, 2004 p.125).

The results in this study also provides support for the notions of curricular alignment, the dominant culture of learning, and teachers’ beliefs systems and their significant role in shaping or mediating the direction and intensity of washback. This leads to a revision of Burrows’ (2004) washback model. Understanding that washback is a broad, multi-faceted and “complex and contentious” matter (Booth, 2012, p. 41), the revised washback model (see Figure 7.8) takes into account test design and its degree of alignment with the focal construct, the prevailing culture of learning, and teachers’ BAK (Woods, 1996) and their “consequent response to change” (e.g., to a newly introduced test or revised testing programme) (Burrows, 2004, p.125).

The revised washback model indicates that the degree of alignment and the washback of the rSECEE is not the only intervening variable in a “quite complicated equation” (Wall, 2000, p.502). Nevertheless, this relationship has possible implications for key stakeholders, such as policy-makers, teachers and students, which are discussed in the next chapter along with the conclusion.

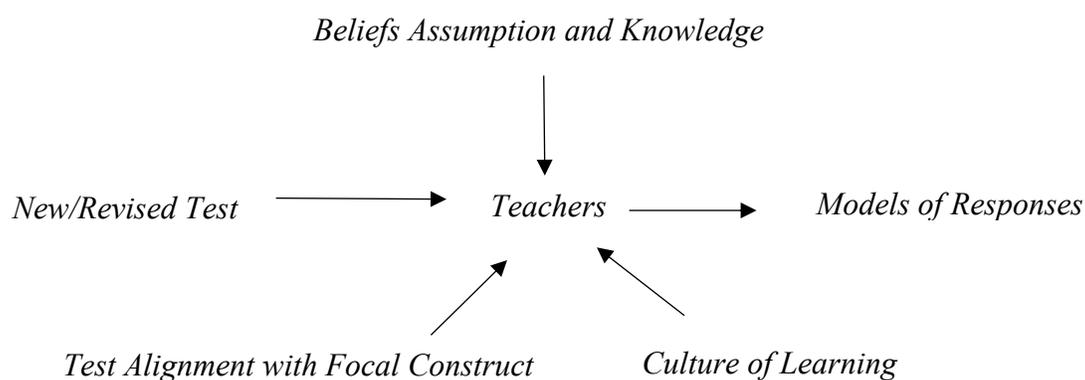


Figure: 7.8: Revised View of Washback  
(Source: Adapted from Burrows 2004, p. 126)

## Chapter XIII

### Conclusion

This chapter provides a summary of the study's main findings, identifies the limitations of the study, highlights its educational implications and provides suggestions for further research.

#### 8.1. Summary of Study

My interest in Libya's national high-stakes testing system is part of a much wider interest in how testing can influence teaching and learning practices at both the school and classroom level i.e., "where the real activity of education occurs" (Chapman & Snyder, 2000, p.458). To address the lack of research in relation to the rSECEE in Libya, the study examined the relationship between the degree of alignment among the components of an educational system (i.e., standards, curriculum, and testing) and the washback of the rSECEE on the classroom. To analyse the relationship, the dissertation first examined the degree of alignment between the rSECEE and Libya's EFL content standards and curriculum (standardised textbooks) (Phase I). Subsequently the dissertation looked into how the washback of the rSECEE operates at the classroom level (Phase II); that is, how a change in a high-stakes national test can influence teachers and teaching, and learners and learning.

To study the relationship between curricular alignment and washback, the study employed a mixed-methods explanatory sequential design (Creswell, 2015) that consisted of two consecutive phases. In Phase I, Webb's (1997, 1999) alignment model was employed to investigate the degree of alignment between the high-stakes test, standards and curriculum. The Webb approach combined both qualitative expert judgment and quantified coding to evaluate the degree of alignment (Flowers et al., 2006). Four criteria were used to judge the degree of alignment:

categorical concurrence; DOK consistency; ROK correspondence; and balance of representation. The researcher acknowledges that “[t]ests are not perfect. Test questions are a sample of possible questions that could be asked in a given area”, besides, a “test score is not an exact measure of a student's knowledge or skills” and that “no single test score can be considered a definitive measure of a student's knowledge” (Heubert & Hauser, 1998, p.3). Nevertheless, it was found that the rSECEE did not meet Webb’s (1997) comprehensive criteria, and there was limited-to-no degree of alignment between the rSECEE and the Libyan EFL content standards. The rSECEE failed to meet the Webb (1997) alignment criteria because it did not: (a) give equal weight to all the Libyan EFL standards; (b) target the full range of depth of knowledge (i.e. from DOK Level 1 to DOK Level 4); (c) cover the full range of Grade 12 EFL content objectives stipulated by the Libyan EFL standards; and (d) represent the target construct (i.e., a case of CUR Messick, 1996). In addition, the rSECEE had a case of CIV (Messick, 1996) because, for example, it included poorly worded test items, questions without a stem, tricky items, and spelling errors. Thus, it can be argued that the rSECEE may not be adequately assessing Grade 12 students’ language competence in the form that is likely to yield accurate information regarding students’ English language skills.

Within the context of misalignment, I looked at how a change in a high-stakes national testing programme and its shortcomings may influence language classrooms. I employed qualitative methods to elicit Phase II data from three Libyan EFL teachers and their students. Participants’ responses to a questionnaire, semi-structured interviews, classroom observations, and focus group interviews contributed in large measure to the Phase II findings. The findings indicated that in a high-stakes testing context, such as Libya, the rSECEE has had mixed effects; it appeared to have negative washback on some teachers and their teaching (content and

practices), while other teachers appeared to have experienced little-to-no evidence of washback. In addition, the results indicated that the rSECEE may have negative effects on learners and their learning. From the findings it can be argued that the washback of the rSECEE is a highly complex matter, and how teachers react to the rSECEE is highly individual. The findings support Cheng (2005), Green (2007) and Watanabe (2004) who argue that the dissonance between the teachers' BAKs and the testing instrument's philosophical underpinnings determines the *direction* and *intensity* of the washback.

Even in the context of a powerful and consequential high-stakes test with high washback potential, it is the teacher who mediates washback –either increasing or diminishing test effects on her students (Fox, 2018 Personal Communication). This finding is consistent with much of the curricular literature which views the teacher as the key stake-holder and the pivotal curriculum maker (e.g., Clandinin, & Connelly, 1992; Fullan, 2001). Following Hargreaves (1989), “change in the curriculum is not effected without some concomitant change in the teacher”, since it is the teacher’s responsibility to convey the curriculum at the classroom level. Thus, “[w]hat the teacher thinks, what the teacher believes, what the teacher assumes all these things have powerful implications for the change process, for the ways in which curriculum policy is translated into curriculum practice” (p.54). The same is true for any newly introduced testing practice.

The collected data from both phases were of value, because they uncovered many areas of rSECEE influence, and thus helped in showing the scope of rSECEE washback. The data also helped me to highlight other important factors within a context of misalignment that can have an influence on what takes place in the Libyan Grade 12 EFL classroom.

In answer to the question: **What is the relationship between the degree of alignment and the washback of the rSECEE?** It can be concluded that national high-stakes tests, such as the rSECEE, as powerfully influential as they are, are not the only mediators in an educational context such as Libya. While the study confirms the complexity and unpredictability of washback, as noted by researchers (e.g., Booth, 2012; Burrows, 2004; Cheng 2005; Green, 2007, 2014; Shohamy et al., 1996; Spratte, 2005; Wall, 2012; Watanabe, 2004), it also suggests that washback complexity can be traced to other contextual factors. Factors, such as the degree of alignment between test-design and the focal construct, prevailing *culture of learning* (Cortazzi & Jin, 1997), and teachers' BAKs (Woods, 1996), play an equally important role in the direction, intensity and *magnitude* (Fox, 2018) of washback. Although the study emphasizes that high-stakes national tests, such as the rSECEE, should be aligned with content standards and the set curriculum, because high-stakes testing inevitably creates incentives for inappropriate methods of test preparation (West, 2010), curricular alignment and the washback of the target test (such as the rSECEE) is not the only intervening variable in a "quite complicated equation" (Wall, 2000, p.502): washback is a multi-faceted, "complex and contentious" matter (Booth, 2012, p. 41). The findings led me to a revision of Burrows' (2004) washback model to include test design and its degree of alignment with the focal construct, the prevailing culture of learning, and teachers' BAKs (Woods, 1996).

Although the findings from this study ought to be interpreted with a degree of caution, they are promising and have possible implications for Libya's stake-holders including policy-makers, test-developers, teachers and students. In addition, the findings can be used to guide efforts to refine and enhance the development of the rSECEE.

## 8.2. Implications for Key Stake-holders in Libya

Considering tests and the testing system play a fundamental role in the Libyan culture of learning and its “long tradition of education that is reflected in individual success in standardised examinations” (Chapman, & Snyder, 2000, p.462), it is legitimate to argue for the promotion of good practice through high-stakes testing such as the rSECEE. This line of argument supports that of Green (2007), Messick (1996), Shohamy (1992) and Wall (2012), and what Thorndike beautifully wrote in 1921 “[s]tudents will work for marks and degrees if we have them. We can have none, or we can have such as worth working for. Either alternative is reasonable, but the second seems preferable” (p.378). Therefore, there may be a better chance of promoting positive washback, if the rSECEE has a better chance of representing the “target skills (whether these are based on a specified curriculum or a target domain), through content, complexity, format, scoring procedures and score interpretations” (Green, 2007, p.13). Huebert and Hauser (1998) further emphasize that teachers tempted to *teach to the test* may be affected by the design of the test. Therefore, it would seem appropriate to prepare students for a test by covering all the test’s requirements which ideally should represent the full domain of content knowledge of the set curriculum. I, therefore, recommend that test developers of the rSECEE cover the full domain of content knowledge and skills operationalised in the Libyan Grade 12 EFL standardised textbook. In addition, if Libyan test developers “sample widely and unpredictably” (Hughes, 2003, p.54), Libyan teachers would likely teach all the curriculum components. In turn, this would likely minimise teachers experiencing “tension between pedagogical and ethical decisions” (Spratt, 2005, p.24), when making instructional decisions at the classroom level. Furthermore, aligning

the rSECEE to the full content domain of knowledge would likely provide students with better opportunities to learn and use English.

Regarding the important role of the traditional culture of learning in Libya and understanding that the Ministry of Education is seeking to develop the quality of the education system by revising the ELT curriculum and its testing systems, it appears that the Ministry is putting the teachers and their students in a challenging situation. Teachers are expected to implement innovative approaches in their classrooms, and to promote and enhance a more productive learning environment for their students, according to the Ministry's EFL standards. However, during the implementation stages teachers have to deal with the prevalent culture of learning that challenges and can impede educational reform and innovation. Consequently, the teachers either have to challenge this culture of learning and adopt the approaches that the curriculum is endorsing or revert to traditional educational approaches. It is clear that one (Teacher C) chose to challenge the existing culture of learning. However, it is also clear that inevitably some teachers, perhaps unconsciously, find themselves teaching in the same way they always have, employing some of the revised curricular practices, but altering them to suit traditional patterns of teaching.

### **8.3. Recommendations**

In the context of misalignment, test developers in Libya need to consider this question: what would the revised Secondary Education Certificate Examination of English (rSECEE) have to look like to encourage Libyan teachers to promote better opportunities for English language learning and use? This study recommends that Libyan rSECEE test developers ought to:

1. Ensure the rSECEE is tailored to the knowledge and skills stated in the Libyan English as a Foreign Language (EFL) content standards;

2. Include challenging content that requires complex demonstrations and application of knowledge and skills from Libyan students. In essence, the rSECEE should not only include test items that are testing depth of knowledge (DOK) Level 1 but all DOK levels, which should ensure they will be taught (Kellaghan & Greaney, 1992);
3. Assess higher-order cognitive skills to ensure they are taught, which as noted by Nuthall and Alton-Lee (1995) should help provide students with a better chance of being more competent and capable of applying the acquired knowledge in real life situations;
4. Incorporate a variety of test formats, including written, oral, aural, because, as argued by Hughes (2003), multiple-choice test items and practices will not provide students with the best possible channels for improving their language abilities. It is also argued by many language test researchers, including Hughes (1998, 2003), that constructed response items such as writing short essays can lead to more positive washback than multiple choice items;
5. Build communicative elements into all parts of the rSECEE
6. Minimise construct under representation (CUR - Messick, 1996) by sampling “widely and unpredictably” (Hughes, 2003, p.54);
7. Minimise construct irrelevant variance (CIV - Messick, 1996) by limiting test item flaws (such as ambiguously worded test items, items testing too much information and not reflecting a specific content, and spelling errors). CIV could also be minimised by test-developers not writing items to deliberately trick students into providing incorrect answers which can encourage both test-wiseness and cheating (Downing, 2002). Both CIV and CUR will be minimised through the implementation of a systematic test validation process, which *tests the test* before a live administration.

8. Achieve transparency in the testing process and products so that all stakeholders including students and parents are able to access information about testing objectives and requirements;
9. Meet with some of the stake-holders in particular Grade 12 teachers, university instructors and programme developers, to develop test specifications which take into account test purpose and the skills Grade 12 EFL students need to have. Hence, the rSECEE would then reflect dialogue among a group of educators, rather than be a top-down dictate. The consultation with stake-holders should help the test developers to decide how: the four language skills are to be weighted; comprehensive will the test be; long will the test be; and the test-scores are to be used; and
10. Design communication networks to “ensure that all stake-holders are kept informed of and are allowed to contribute to new developments” (Wall, 1996, p.352).

By incorporating the above recommendations, there should be an improved chance of increasing the degree of alignment between the Libyan EFL content standards and the rSECEE. If an adequate degree of alignment was to exist between the Libyan EFL content standards and the rSECEE indicating that the majority of the Libyan EFL standards are assessed and that there is a balance between standards and the rSECEE assessment items, a strong link is created between the three components of the Libyan education system. With this strong link, teachers should be motivated to teach all standards, and thus, Libyan EFL classroom teaching would be more closely aligned to both the Libyan EFL standards and the rSECEE. In addition, with the high degree of alignment in place, the rSECEE would assess the students’ attainment of the expected knowledge and standards. Consequently, the students would have a better opportunity to attain all the standards, and, as noted by Linn (1994), their learning should be at a higher level.

In addition, the results of the rSECEE would provide valid information on how well the students had attained the set standards, and thus have a lower washback *magnitude*.

Moreover, the study considers the central and influential role of the teacher in the relationship between curricular alignment and washback because, as emphasized by both Atkin (1992) and Whitehead (1989), teachers attempt to make changes once they recognise the presence of a gap or inconsistency between their objectives and principles and their current practices. Furthermore, it is believed by Hashweh (2003, p.421) that teachers experience accommodative changes when they: (a) are internally motivated to learn; (b) become aware of their implicit ideas and practices and critically examine them; and (c) construct alternative knowledge, beliefs, and practices. Importantly, these conditions can only take place within a social environment that is characterized by cooperation, trust, reflection and thoughtfulness (Hashweh, 2003). Under these conditions, Libyan Grade 12 EFL teachers would likely be willing to change, because “[p]eople are intrigued when they see good things happening in the lives of individuals, families and organizations... And their immediate request is very revealing of their basic paradigm. ‘How do you do it?’ Teach me the techniques” (Covey, 1989, p.40). Accordingly, as suggested by Cheng et al. (2015), teachers need to be informed about how testing has the potential to shape what they teach, and testing may influence the practices to include test preparation material and activities through professional development programmes. Cheng et al. (2015) further recommend that teacher education programmes include a focus on the assessment and evaluation of learning in general and on high-stakes tests and washback in particular, as this would likely help teachers recognise the variety of interrelated factors within the washback phenomena. Similarly, as recommended by Kellaghan and Greaney (1992),

professional networks within Libya ought to be “developed to initiate exchange programmes and to share common interests and concerns” (p.3).

Adapting Chapman and Snyder’s (2000) model of linkage between high-stakes testing and instructional practices, Figure 8.1 illustrates the logic behind how the recommendations discussed above can connect the possible uses of high stakes testing to improve instructional practice. Moreover, the study’s recommendations also cover the Libyan culture of learning. If the Libyan Ministry of Education wants to initiate a reform policy at the secondary level, it ought to deal with the problems arising from the traditional culture of learning, such as the prominence of memorisation, as early as Grade 1. As reported by Abdulhamid (2011), the prominence of memorisation skills within the Libyan education system impedes innovation and reform. It can be argued that memorisation as part of the Libyan culture of learning shapes how any curriculum is delivered, how testing is developed, and how learning is translated to students. In addition, “[i]mages of classrooms which teachers have constructed from years of experience in schools, both as students and teachers, cannot be changed with words alone” (Briscoe, 1991, p.198).

Therefore, as argued by Abdulhamid (2011), the Ministry of Education’s reform policy ought to re-consider future Libyan classroom pedagogy (from Grade 1 to university-level education), so that rote learning and memorisation have a lower prominence in classroom practices and tests. Instead, it should promote opportunities for students to become analytical, critical, and autonomous learners, because a “teacher who is attempting to teach without inspiring the pupil with a desire to learn, is hammering on cold iron” (Horace Mann, 1796-1859, cited in Pychyl, 2011). Based on the teachers’ and students’ accounts and my experience it can be argued that the proposed reform for secondary level institutions ought to include: increasing resources and facilities; increasing training for teachers; and lowering the teaching load.

*Possible Uses of Testing*

*Linkage to Instructional Practices*

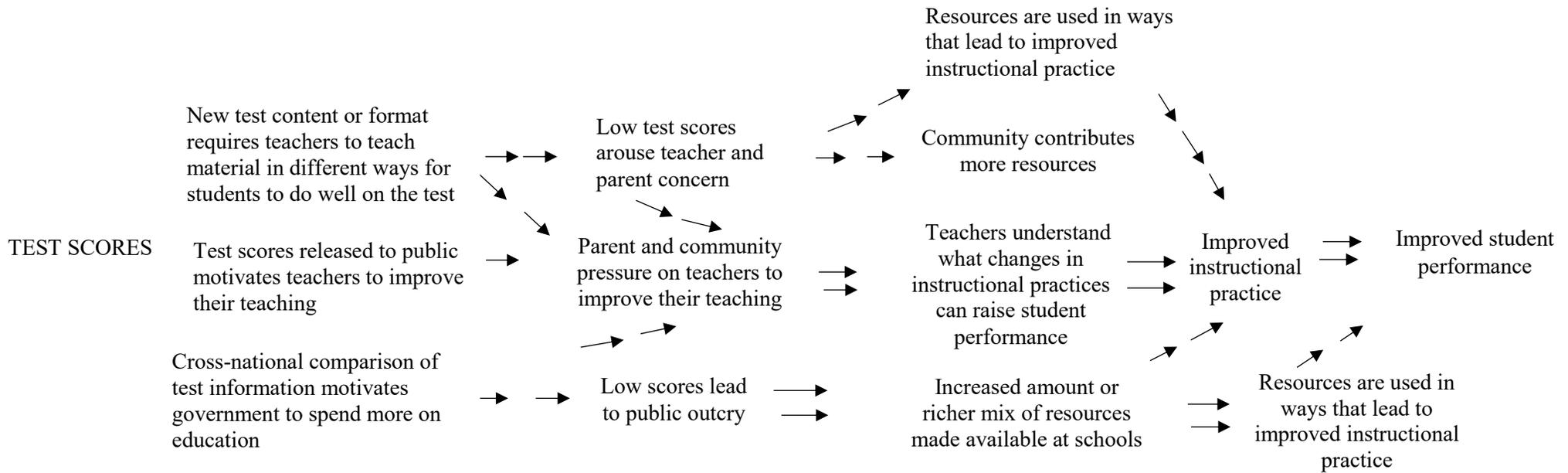


Figure 8.1: A Model of Linkage between the High-Stakes Test and Instructional Practices

Source: Adapted from Chapman and Snyder (2000, p.466).

Human, material, and financial resource support as suggested in Figure 8.1 is essential for the successful implementation of any curricular innovation, including testing (Li, 1998). By considering this along with the other suggested elements for reform, the Ministry of Education could take steps towards filling the gaps within the Libyan education system. In this way, *change* in materials, methods, and testing would not simply occur at the surface level, but would represent a deep and enriching increase in understanding and knowledge that encourages life-long learning for both teachers and students.

#### **8.4 Limitations and Future Research**

Although I tried to elicit as much rigorous data as possible and triangulated the analysis, there are factors that could have affected the findings. This study did not have a baseline study that could have uncovered other factors that may affect the relationship between alignment and washback. A second limitation was that the content standards were written so generally, which, in turn, required so many content objectives to be incorporated, and thus made determining a match even more difficult. Another related problem is what La Marca et al. (2000) noted, that testing items may measure multiple content standards. For these reasons and, as emphasized by Ananda (2003), perfect alignment can never be expected.

Moreover, the results of this research were limited by the lack of access to the test specifications, i.e., the test blueprints. The researcher was unable to obtain more information that documents the test design and test and item development. Having such access could have provided me with the opportunity to evaluate how well the blueprints for the rSECEE represented the Ministry's goals for learning. The study may also be limited in that it did not

evaluate the development process of the rSECEE. As recommended by Forte (2016), an alignment evaluation programme needs to study the whole system, and not just a single test. She further notes that a comprehensive alignment study needs to assess both the development process of the target assessment and the outcome of the development process. The data obtained can yield useful information for guiding and improving an education system. The evaluation of the development process could have helped me to collect evidence about how reasonable and clear the development process is in terms of yielding reliable and aligned assessment instruments. This is because “[i]t is far better to discover weaknesses in the foundation before the rest of the house is built” (Forte, 2016, p.18).

Meanwhile, future research could focus on collecting data from the rSECEE development process and, in accordance with Forte (2016), the review process could address the following research questions:

- How was the SECEE developed to reflect the Libyan EFL standards?
- How were the test specifications developed to reflect the target standards?

The thoroughness of this study could be increased by the collection of additional documentation, such as those verifying how the standards, test blueprints, and test items, were developed. In addition, reports that document how item writers and test developers were trained in the development process and reports that document pilot testing results would also be advantageous for research rigorosity.

I would like to end this dissertation with a note that other high stakes tests such as TOEFL have benefitted from critical discussions related to its test design and impact, and such discussions have essentially lead to its development and improvement (Baird, Andrich, Hopfenbeck, & Stobart, 2017). To help develop and improve the rSECEE, its users could benefit

from such discussions. In my view, when done systematically, large-scale testing has the potential to improve teaching and to create opportunities for learning (Huebert & Hauser, 1998). Although tests may be used improperly, this “should not discourage policymakers, teachers, and parents. Rather, it should motivate action to ensure that educational tests are used fairly and effectively” (Huebert & Hauser, 1998, p.9). This study may hopefully contribute to that essential work.

## References

- Abdelfattah, F. (2010). The relationship between motivation and achievement in low-stakes examinations. *Social Behavior and Personality: an international journal*, 38(2), 159-167.
- Abdulhamid, N. (2011). What is the impact of the Libyan Study Abroad Scholarship Programme on returning university-level English teachers?. Unpublished PhD thesis Carleton University.
- About Jaafar, A. E. (2003) The EFA 2000 Assessment: Country Reports, Libyan Jamahiriya. Retrieved from [http://www2.unesco.org/wef/countryreports/libya/rapport\\_1.html](http://www2.unesco.org/wef/countryreports/libya/rapport_1.html)
- Abubaker, F. M. H. (2017). *The road to possibilities: a conceptual model for a program to develop the creative imagination in reading and responding to literary fiction (short stories) in Libyan English as a Foreign Language (EFL) university classrooms*. Unpublished PhD thesis, University of Glasgow.
- Agee, J. (2004). Negotiating a Teaching Identity: An African American Teacher's Struggle to Teach in Test-Driven Contexts. *Teachers College Record*, 106(4), 747-774.
- Airasian, P. W. (1987). State mandated testing and educational reform: Context and consequences. *American Journal of Education*, 393-412.
- Aldabbus, S. (2008). An investigation into the impact of language games on classroom interaction and pupil learning in Libyan EFL primary classrooms. (Unpublished Doctoral dissertation). Newcastle University, United Kingdom.

- Alderson, J Charles (1986) Innovations in language testing? In Portal, M (ed) *Innovation in Language Testing*. (pp. 93-105). Windsor: NI:RE-Nelson.
- Alderson, J. C. (2004). Forward. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in language testing: Research context and methods* (pp. ix-xii). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.
- Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics*, 14(2), 115–129.
- Alderson, J.C. & Hamp-Lyons, L. (1996). TOEFL preparation courses: a study of washback. *Language Testing*, 13 (3), 280-297.
- Al-Hamdan, Z., & Anthony, D. (2010). Deciding on a mixed-methods design in a doctoral study. *Nurse researcher*, 18(1), 45-56.
- Alhmali, R. (2007) *Students Attitudes in the Context of the Curriculum in Libyan Education in Middle and High Schools*, unpublished PhD thesis. University of Glasgow.
- Ali, M. A. A. (2008). *The Oral Error Correction Techniques Used by Libyan Secondary School Teachers of English* (Doctoral dissertation, University of Sunderland).
- Allwright, D. & Bailey, K. (1991). *Focus on the language classroom*. New York: Cambridge University Press.
- Aloreibi, A., & Carey, M. D. (2017). English language teaching in Libya after Gaddafi. In *English Language Education Policy in the Middle East and North Africa* (pp. 93-114). Springer International Publishing.

- Altaieb, S. (2013). *Teachers' Perception of the English language Curriculum in Libyan Public Schools: An investigation and assessment of implementation process of English curriculum in Libyan public high schools* (Unpublished Doctoral Dissertation, University of Denver).
- American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (1999). *Standards for educational and psychological testing*. Amer Educational Research Assn.
- Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing & student learning. *Education policy analysis archives, 10*, 18.
- Amrein, A.L., & Berliner, D.C. (2002a). An Analysis of Some Unintended and Negative Consequences of High-Stakes Testing. Education Policy Research Unit, Arizona State University, Tempe, AZ. Retrieved from <http://epsl.asu.edu/epru/documents/EPSSL-0211-125-EPRU-exec.pdf>.
- Amrein, A. L., & Berliner, D. C. (2002b). High-Stakes Testing & Student Learning. *Education Policy Analysis Archives, 10*(18), n18.
- Anagnostopoulos, D. (2003). Testing and student engagement with literature in urban classrooms: A multi-layered perspective. *Research in the Teaching of English, 177-212*.
- Ananda, S. (2003). Achieving alignment. *Leadership, 55*(1), 18-22.
- Anderson, G. E., Whipple, A. D., & Jimerson, S. R. (2002). Grade retention: Achievement and mental health outcomes. *National Association of School Psychologists, 1-4*.

- Anderson, J. (1993). Is a Communicative Approach Practical for Teaching in China? Pros and Cons. *System*, 21 (4). 471-480.
- Anderson, L. W. (2002). Curricular alignment: A re-examination. *Theory into practice*, 41(4), 255-260.
- Andrews, S. (1994). The washback effect of examinations: Its impact upon curriculum innovation in English language teaching. *Curriculum Forum 1* (4).
- Andrews, S. (1995). Washback or washout? The relationship between examination reform and curriculum innovation. In D. Nunan, V. Berry & R. Berry (Eds.), *Bringing about change in language education* (pp. 67-81). Hong Kong: University of Hong Kong.
- Andrews, S. J. (2004). Washback and curriculum innovation In Y. J. W. L. Cheng, with A. Curtis (Ed.), *Washback in language testing: Research contexts and methods* (pp. 37–52). Mahwah, NJ: Lawrence Erlbaum.
- Andrews, S., Fullilove, J., & Wong, Y. (2002). Targeting washback—a case-study. *System*, 30(2), 207-223.
- Apple, M. W. (1995). *Education and power* (2<sup>nd</sup> ed.). New York: Routledge.
- Asker, A. (2012). Future self-guides and language learning engagement of English-major secondary school students in Libya: Understanding the interplay between possible selves and the L2 learning situation (Doctoral dissertation, University of Birmingham).
- Atkin, J. M. (1992). Teaching as research: An essay. *Teaching and Teacher Education*, 8(4), 381–390.

- Au, W. (2007). High-stakes testing and curricular control: A qualitative meta-synthesis. *Educational Researcher*, 36(5), 258-267.
- Au, W. W. (2008). Devising inequality: a Bernsteinian analysis of high-stakes testing and social reproduction in education. *British Journal of Sociology of Education*, 29(6), 639-651.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (2010). *Language Assessment in Practice*. Oxford: Oxford University Press.
- Bachman, L.F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bagigni, A. (2016). *The Role of English in Libya and its Implications for Syllabus Design in Libyan Higher Education* (Doctoral Dissertation, University of Huddersfield).
- Bailey, K. M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing*, 13, 257-279.
- Bailey, K. M. (1999a). *Washback in language testing*. TOEFL Monograph Series, Ms. 15. Princeton, NJ: Educational Testing Service.
- Bailey, K. M. (1999). *Washback in language testing*. TOEFL Monograph Series, Ms. 15. Princeton, NJ: Educational Testing Service.
- Baird, J. A., Andrich, D., Hopfenbeck, T. N., & Stobart, G. (2017). Assessment and learning: Fields apart? *Assessment in Education: Principles, Policy & Practice*, 24(3), 317-350.

- Baker, B. A. (2010). *In the service of the stakeholder: A critical, mixed methods program of research in high-stakes language assessment* (doctoral dissertation).
- Baker, E. (1991). Alternative assessment and national policy. Paper presented at the National Research Symposium on Limited English Proficient Students' Issues: Focus on Evaluation and Measurement, Washington, DC.
- Baker, E. L. (2004). *Aligning curriculum, standards, and assessments: Fulfilling the promise of school reform* (CSE report 645). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.
- Baker, E. L. (2005). Aligning curriculum, standards, and assessments: Fulfilling the promise of school reform. *Measurement and research in the accountability era*, 315-335.
- Barr, J. J. (2016). Developing a Positive Classroom Climate. IDEA Paper # 61. *IDEA Center, Inc.*
- Baskarada, S. (2014). Qualitative case study guidelines. *Qualitative Report*, 19(1), 1-25.
- Baxter, P., & Jack, S. (2008). Qualitative case study methodology: Study design and implementation for novice researchers. *The qualitative report*, 13(4), 544-559.
- Bazeley, P. (2012). *Qualitative data analysis: Practical strategies*. Thousand Oaks, CA: Sage Publications, Inc.
- Beauchamp, G. A. (1982). Curriculum theory: Meaning, development, and use. *Theory into practice*, 21(1), 23-27.
- Becker, C., & Roos, J. (2016). An approach to creative speaking activities in the young learners' classroom. *Education Inquiry*, 7(1), 9-26.

- Bertenthal, M. W., & Wilson, M. R. (Eds.). (2005). *Systems for state science assessment*. National Academies Press.
- Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice*, 22(3), 21-29.
- Biggs, J. (2003). Aligning teaching and assessing to course objectives. *Teaching and Learning in Higher Education: New Trends and Innovations*, 2, 13-17.
- Biggs, J., & Tang, C. (2003). *Teaching for Quality Learning at University*, Society for Research into Higher Education and Open University Press. *New edition*.
- Biggs, J., & Tang, C. (2011). *Teaching for quality learning at university: What the student does*. McGraw-Hill Education (UK).
- Blackwell, S. (2003) Saving the King: Anglo-American strategy and British counter supervision operations in Libya, 1953-59. *Middle Eastern Studies*, 39, 1-18.
- Blazer, C. (2011). Unintended Consequences of High-Stakes Testing. Information Capsule. Volume 1008. *Research Services, Miami-Dade County Public Schools*.
- Bloom, B. S., Madaus, G. F., & Hastings, J. T. (1981). *Evaluation to Improve Learning*. New York: McGraw-Hill.
- Booher-Jennings, J. (2008). Learning to label: Socialisation, gender, and the hidden curriculum of high-stakes testing. *British Journal of Sociology of Education*, 29(2), 149-160.
- Booth, D. K. (2012). *Exploring the Washback of the TOEIC in South Korea: A sociocultural perspective on student test activity* (Doctoral dissertation, ResearchSpace@ Auckland).

- Borg, M. (2001). Teachers' beliefs. *ELT Journal*, 55(2), 186-188.
- Bransford, J. D., & Schwartz, D. L. (1999). Chapter 3: Rethinking transfer: A simple proposal with multiple implications. *Review of research in education*, 24(1), 61-100.
- Braun, H. (2004). Reconsidering the Impact of High-stakes Testing. *Education Policy Analysis Archives*, 12(1), 01.
- Breen, M., Hird, B., Milton, Oliver, R., & Thwaite, A. (2001). Making sense of language teaching: Teachers! Principles and classroom practices. *Applied Linguistics*, 22(4), 470-501.
- Brimijoin, K. (2005). Differentiation and high-stakes testing: An oxy- moron? Theory Into Practice, 44(3), 254-261.
- Brindley, G. (Ed.). (2000). *Studies in immigrant English language assessment* (Vol. 1). Sydney, Australia: National Centre for English Language Teaching and Research, Macquarie University.
- Briscoe, C. (1991). The dynamic interactions among beliefs, role metaphors, and teaching practices: A case study of teacher change. *Science Education*, 75(2), 185- 199.
- Broadfoot, P. M. (2005). Dark alleys and blind bends: Testing the language of learning. *Language Testing*, 22(2), 123-141.
- Brown, H. D. (2000). *Teaching by principles: An interactive approach to language pedagogy*. New York: Pearson.
- Brown, J. D. (2001). *Using surveys in language programs*. Cambridge, England: Cambridge University Press.
- Brualdi, A. (1999). Traditional and Modern Concepts of Validity. ERIC/AE Digest.

- Buckendahl, C. W., Plake, B. S., Impara, J. C., & Irwin, P. M. (2000, April). Alignment of standardized achievement tests to state content standards: A comparison of publishers' and teachers' perspectives. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.
- Burns, R.B. (1997). *Introduction to research methods*. (3rd ed.) Australia: Longman.
- Burrows, C. J. (1998). Searching for washback: an investigation of the impact on teachers of the implementation into the Adult Migrant English Program of the assessment of the Certificate in Spoken and Written English. (Doctoral dissertation, Macquarie University).
- Burrows, C. (2004). Washback in classroom-based assessment: A study of the washback effect in the Australian adult migrant English program. *Washback in language testing: Research contexts and methods*, 113-128.
- Case, B. J., Jorgensen, M. A., & Zucker, S. (2004). Alignment in educational assessment. Retrieved on September, 5, 2016, from [https://images.pearsonassessments.com/images/tmrs/tmrs\\_rg/AlignEdAss.pdf?WT.mc\\_id=TMRS\\_Alignment\\_in\\_Educational\\_Assessment](https://images.pearsonassessments.com/images/tmrs/tmrs_rg/AlignEdAss.pdf?WT.mc_id=TMRS_Alignment_in_Educational_Assessment)
- Chang, S. C. (2011). A contrastive study of grammar translation method and communicative approach in teaching English grammar. *English Language Teaching*, 4(2), 13.
- Chapman, D. W., & Snyder, C. W. (2000). Can high stakes national testing improve instruction: re-examining conventional wisdom. *International Journal of Educational Development*, 20(6), 457-474.

- Charamaz, K. (2006). *Constructing grounded theory: A practical Guide through qualitative research analysis*. London: Sage.
- Chen, L. (2002). *Washback of a public examination: Impact of the Basic Competency Test on Taiwan junior high school English teaching*, presented in Edward F. Hayes Graduate Research Forum and paper appeared in Edward F. Hayes Graduate Research Proceedings, 2, 91-99.
- Cheng, L. (1997). How does washback influence teaching? Implications for Hong Kong. *Language and education*, 11(1), 38-54.
- Cheng, L. (1998). Impact of a public English examination change on students' perceptions and attitudes toward their English learning. *Studies in Educational Evaluation*, 24(3), 279-301.
- Cheng, L. (1999). Changing assessment: Washback on teacher perceptions and actions. *Teaching and teacher education*, 15(3), 253-271.
- Cheng, L. (2001). Washback studies: Methodological considerations. In *Curriculum Forum* (Vol. 10, No. 2, pp. 17-32).
- Cheng, L. (2003). Looking at the impact of a public examination change on secondary classroom teaching: A Hong Kong case study. *Journal of Classroom Interaction*, 38 (1), 1-10
- Cheng, L. (2004). The washback effect of a public examination change on teachers' perceptions toward their classroom teaching. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 147-170). Mahwah, NJ: Lawrence Erlbaum.

- Cheng, L. (2005). *Changing language teaching through language testing: A washback study* (Vol. 21). Cambridge University Press.
- Cheng, L. (2008). The key to success: English language testing in China. *Language Testing*, 25(1), 15-37.
- Cheng, L. (2008). Washback, impact and consequences. In Shohamy, E. & Hornberger, N. H. (Eds.), *Encyclopedia of language and education. Volume 7: Language testing and assessment*, 2nd edn. (pp. 349–364). New York: Springer Science and Business Media LLC.
- Cheng, L. (2014). Consequences, impact, and washback. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1130–1146). New York, NY: John Wiley & Sons. doi:10.1002/9781118411360.wbcla071
- Cheng, L., & Couture, J. C. (2000). Teachers' Work in the Global Culture of Performance. *Alberta Journal of Educational Research*, 46(1), 65-74.
- Cheng, L., & Curtis, A. (2004). Washback or backwash: A review of the impact of testing on teaching and learning. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 3-18). Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Cheng, L., & Fox, J. (2017). *Assessment in the language classroom: Teachers supporting student learning*. London: Palgrave.
- Cheng, L., Andrews, S. & Yu, Y. (2011). Impact and consequences of school-based assessment in Hong Kong: Views from students and their parents. *Language Testing*, 28(2), 221-250.

- Cheng, L., Myles, J. & Curtis, A. (2004). Targeting language support for non-native English-speaking graduate students at a Canadian university. *TESL Canada Journal*, 21(2), 50-71.
- Cheng, L., Sun, Y., & Ma, J. (2015). Review of washback research literature within Kane's argument-based validation framework. *Language Teaching*, 48(04), 436-470.
- Cheng, L., Watanabe, Y., & Curtis, A. (Eds.). (2004). *Washback in language testing: Research contexts and methods*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Cimbricz, S. (2002). State-mandated testing and teachers' beliefs and practice. *education policy analysis archives*, 10, 2.
- Clandinin, D. J., & Connelly, F. M. (1992). Teacher as curriculum maker. In P. Jackson (Ed.), *Handbook of research on curriculum* (pp. 363-401). New York: Macmillan.
- Clark, E. V., & Hecht, B. F. (1983). Comprehension, production, and language acquisition. *Annual review of psychology*, 34(1), 325-349.
- Cohen, A. D. (1994). *Assessing language ability in the classroom* (2nd ed.). New York: Heinle & Heinle.
- Cohen, L., & Manion, L. (1989). *Research methods in education*. London: Routledge.
- Cohen, S.A. (1987). Instructional alignment: Searching for a magic bullet. *Educational*
- Cole, J. S., Bergin, D. A., & Whittaker, T. A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology*, 33(4), 609-624.
- Cole, N. S. (1999). Determining what is to be taught: The role of assessment. *ENC Focus: Assessment that Informs Practice*, 7(2), 34.
- Coleman, H. (1996). *Society and the Language Classroom*. Cambridge: Cambridge.

- Connelly, F. M., Clandinin, D. J., & Fullan, M. (1993). *Teacher education: Links between personal and professional knowledge*. Toronto, ON: Social Sciences and Humanities Research Council, Joint Centre for Teacher Development. *Ontario Institute for Studies in Education, and University of Toronto*.
- Cook-Sather, A. (2010). Students as learners and teachers: Taking responsibility, transforming education, and redefining accountability. *Curriculum Inquiry*, 40(4), 555-575.
- Corbet, H.D., Wilson, B., 1991. *Testing, Rebellion and Reform*. Ablex, Norwood, NJ.
- Cortazzi, M. & Jin, L. (1997). Culture of learning: Language classrooms in China. In H. Coleman (Ed.), *Society and the language classroom*, pp. 169-206. Cambridge: Cambridge University Press.
- Country Studies (1987) Libya: The Fourth Shore [online]. Library of Congress. Retrieved from [http://lcweb2.loc.gov/cgi-bin/query/r?frd/cstdy:@field\(DOCID+ly0031](http://lcweb2.loc.gov/cgi-bin/query/r?frd/cstdy:@field(DOCID+ly0031)
- Covey, S. R. (1989). *The 7 habits of highly effective people: Powerful lessons in personal change* (Vol. 247). New York: Simon & Schuster.
- Crabtree, B. F. & Miller, W. L. (1999). *Doing qualitative research*. 2nd edit., Thousand Oaks, CA: Sage Publications.
- Creswell, J. W. (2015). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.
- Creswell, J. W., & Miller, D. L. (2000). Determining validity in qualitative inquiry. *Theory into practice*, 39(3), 124-130.

- Creswell, J. W., & Plano Clark, V. L. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage
- Creswell, J. W., Klassen, A. C., Plano Clark, V. L., & Smith, K. C. (2011). Best practices for mixed methods research in the health sciences. *Bethesda (Maryland): National Institutes of Health*, 2094-2103.
- Creswell, J.W. (2003). *Research design: Qualitative, quantitative, and mixed methods approaches*. (2nd ed.) Thousand Oaks: Sage.
- Creswell, J. And Plano Clark, V. (2011). *Designing and Conducting Mixed Methods Research*. Thousand Oaks, CA: Sage.
- Crocker, L. M., Miller, M. D., & Franks, E. A. (1989). Quantitative methods for assessing the fit between test and curriculum. *Applied Measurement in Education*, 2, 179–194.
- Dagget, W. R. (2000). Mowing from standards to instructional practice. *NASSP Bulletin*, 84(66), 66-72.
- Darling-Hammond, L. (2000). How teacher education matters. *Journal of Teacher Education*, 51, 166–173.
- Darlington, Y., & Scott, D. (2002). *Qualitative research in practice: Stories from the field*. Buckingham: Open University Press.
- Davidson, F. & Lynch, B.K. (2002) *Test craft. A teacher's guide to writing and using language test specifications*. New Haven and London: Yale University Press.

- De Lano, L., Riley, L., Crookes, G. (1994). The meaning of innovation for ESL teachers. *System* 22 (4), 487–496.
- Debray, E., Parson, G., & Avila, S. (2003). Internal alignment and external pressure. In M. Carnoy, R. Elmore, & L. S. Siskin (Eds.), *The new accountability: High schools and high-stakes testing* (pp. 55–85). New York: RoutledgeFalmer.
- Deeb, K., M., & Deeb, I., M. (1982). *Libya since the Revolution: Aspects of Social and Political Development*. New York: Praeger Publishers,
- Dennen, V. P., & Burner, K. J. (2007). The cognitive apprenticeship model in educational practice. *Handbook of research on educational communications and technology*, 425-439.
- Denscombe, M. (2010). *The good research guide: For small-scale social research projects (Open UP Study Skills)*. McGraw-Hill.
- Denzin, N. K., & Lincoln, Y. S. (2000). *The SAGE handbook of qualitative research*. Sage.
- Dewey, J. (1916). *Democracy and education*. New York: The Free Press.
- Dewey, J. (1933). How we think. (rev. ed.). Boston: D.C. Heath*
- Dillon, S. (2006). Schools cut back subjects to push reading and math. *The New York Times*, p. 1.
- Dodd, D, K., & Leal, L. (2002). Answer justification: removing the “trick” from multiple-choice questions. In R.A. Griggs (Ed.), *Handbook for teaching introductory psychology* (Vol. 3, pp. 99-100). Mahwah, NJ: Lawrence Erlbaum Associates.
- Donaghue, H. (2003). An instrument to elicit teachers' beliefs and assumptions. *ELT journal*, 57(4), 344-351.

- Donn, G. (1994). Feminist approaches and the curriculum. In T. Husén & T. N. Postlethwaite (Eds.), *The international encyclopedia of education*. Volume 3. (pp. 2287-2292). Oxford: Pergamon.
- Downing, S. M. (2002). Threats to the validity of locally developed multiple-choice tests in medical education: construct-irrelevant variance and construct underrepresentation. *Advances in Health Sciences Education*, 7(3), 235-241.
- Downing, S. M., & Haladyna, T. M. (2004). Validity threats: overcoming interference with proposed interpretations of assessment data. *Medical Education*, 38(3), 327-333.
- Dughri, M. (1980). Human Resources Development and Educational Policy in Libya. Doctoral Dissertation. Retrieved from ProQuest Dissertations and Theses 1980. (AAT 8018296)
- El Abbar, M. (2016). A Lesson Study of Internet Usage to Enhance the Development of English Language Teaching in a Libyan University (Doctoral dissertation, University of East Anglia).
- Elabbar, A. A. (2011). *An investigation of influences affecting Libyan English as Foreign Language University Teachers (LEFLUTs), teaching approaches in the language classrooms* (Doctoral dissertation, University of Glasgow).
- Elder, C., & Wigglesworth, G. (1996). Perspectives on the testing cycle: Setting the scene. *Australian Review of Applied Linguistics, Series S(13)*, 1-12.
- Elliott, S. N., Braden, J. B., & White, J. L. (2001). *Assessing one and all: Educational accountability for students with disabilities*. Arlington, VA: Council for Exceptional Children.

- Elmabruk, R. (2008). Using the Internet to Support Libyan In-service EFL Teachers' Professional Development. Doctoral Dissertation. University of Nottingham, United Kingdom. Retrieved from [http://etheses.nottingham.ac.uk/1038/1/Elmabruk\\_\(2008\)\\_PhD\\_Thesis.pdf](http://etheses.nottingham.ac.uk/1038/1/Elmabruk_(2008)_PhD_Thesis.pdf)
- Elmore, R. F. (1996). Getting to scale with good educational practice. *Harvard Educational Review*, 66 (1), 1 - 26.
- English, F. W. (1992). *Deciding what to teach and test: Developing, aligning, and auditing the curriculum*. Newbury Park, CA: Corwin Press.
- Erfani, S. (2013). Comparative washback study of IELTS and TOEFL iBT on teaching and learning activities in preparation courses in the Iranian context. *English Language Teaching*, 5(8), 185-195.
- Fairbairn, S. & Fox, J. (2009). Inclusive achievement testing for linguistically and culturally diverse test takers: Essential considerations for test developers and decision makers. *Educational Measurement: Issues and Practice*, 28(1), 10–24.
- Farenga, S. J., Joyce, B. A., & Ness, D. (2002). Reaching the zone of optimal learning: The alignment of curriculum, instruction, and assessment. In R. W. Bybee (Ed.). *Learning science and the science of learning*, (pp.51-62). Arlington: NSTA press.
- Farhady, H. (2018). History of Language Testing and Assessment. In *The TESOL Encyclopedia of English Language Teaching*.
- Ferman, I. (2004). The washback of an EFL national oral matriculation test to teaching and learning. *Washback in language testing: Research contexts and methods*, 191-210.

- Fickel, L. H. (2006). Paradox of practice: Expanding and contracting curriculum in a high-stakes climate. In S. G. Grant (Ed.), *Measuring history: Cases of state-level testing across the United States* (pp. 75–103). Greenwich, CT: Information Age Publishing.
- Firestone, W.A., Goertz, M.E., & Natriello, G.J. (1997). *From cashbox to classroom: The struggle for fiscal reform and educational change in New Jersey*. New York: Teachers College Press.
- FitzPatrick, B., Hawboldt, J., Doyle, D., & Genge, T. (2015). Alignment of learning objectives and assessments in therapeutics courses to foster higher-order thinking. *American journal of pharmaceutical education*, 79(1), 10.
- Fives, H., & DiDonato-Barnes, N. (2013). Classroom test construction: The power of a table of specifications. *Practical Assessment, Research & Evaluation*, 18(3), 2.
- Flores, B. B., & Clark, E. R. (2003). Texas voices speak out about high-stakes testing: Preservice teachers, teachers, and students. *Current Issues in Education*, 6(3). Retrieved from <http://cie.ed.asu.edu/volume6/number3/>
- Flowers, C., Browder, D., & Ahlgrim-Delzell, L. (2006). An analysis of three states' alignment between language arts and mathematics standards and alternate assessments. *Exceptional Children*, 72(2), 201-215.
- Forte, E. (2016). *Evaluating alignment in large-scale standards-based assessment systems*. Washington, DC: Technical Issues in Large Scale Assessment SCASS of CCSSO.
- Fox, J. (2004). Curriculum design: Does it make a difference? *Contact, Special Research Symposium Issue*, 30(2), pp. 1–5

- Fox, J., & Cheng, L. (2007). Did we take the same test? Differing accounts of the Ontario Secondary School Literacy Test by first and second language test-takers. *Assessment in Education: Principles, Policy and Practice*, 14(1), 9–26.
- Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational researcher*, 18(9), 27-32.
- Frederiksen, N. (1984). The influence of testing on teaching and learning. In J. Herman (Ed.), *Wagging the dog, carting the horse: Testing and improving schools*. Summary of conference proceedings, Research into Practice Project. Los Angeles: University of California at Los Angeles, Center for the Study of Evaluation.
- Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist*, 39(3), 193–202.
- Freeman, D., & Richards, I. (1996). *Teacher Learning in Language Teaching*. Cambridge: Cambridge University Press.
- Fulcher, G. (1997). An English language placement test: issues in reliability and validity. *Language testing*, 14(2), 113-139.
- Fulcher, G. (2011) 'Cheating gives lies to our test dependence: Policymakers are using language tests to carry a larger social burden than they can reasonably bear'. *The guardian Weekly*, Tuesday 11th October 2011, *Learning English* (4), <http://www.guardian.co.uk/education/2011/oct/11/why-more-language-testcheating?INTCMP=SRCH>.

- Fullan, M. (2001). *The New Meaning of Educational Change* (3rd ed.). London: Teachers College Press.
- Fullilove, J. (1992) The tail that wags. *Institute of Language in Education Journal* 9, 131– 47.
- Futrell, M. H., & Rotberg, I. C. (2002). Predictable casualties. *Education Week*, 22(5), 34, 48.
- Gao, Y. (2003). Need for Scientific Evidence in Foreign Language Teaching and Learning Reforms. *Foreign Language Teaching and Research*, 3, 222-223.
- Gay, L. R., & Airasian, P. (2000). *Educational research: Competencies for analysis and applications*. Upper Saddle River, NJ: Merrill.
- Gayler, K., Chudowsky, N., Kober, N., & Hamilton, M. (2003). State High School Exit Exams Put to the Test. Center on Education Policy, Washington, DC.
- Gerring, J. (2004). What is a case study and what is it good for? *American Political Science*
- Gerwin, D., & Visone, F. (2006). The Freedom to teach: Contrasting History teaching in elective and state–tested courses. *Theory & Research in Social Education*, 34(2), 259-282.
- Ghuma, M.A. (2011). *The Transferability of Reading Strategies between L1 (Arabic) and L2 (English)*. Unpublished PhD thesis, University of Durham.
- Gibbs, A. (1997). Focus groups. *Social research update*, 19(8), 1-8.
- Gipps, C., McCallum, B., & Brown, M. (1996). Models of teacher assessment among primary school teachers in England. *The Curriculum Journal*, 7(2), 167-183.
- Glover, P. (2014). Do language examinations influence how teachers teach. *International Online Journal of Education and Teaching/ISSN: 2148-225X*, 1(3).

- Golafshani, N. (2003). Understanding reliability and validity in qualitative research. *The qualitative report*, 8(4), 597-606.
- Gorard, G. (2004). *Combining methods in educational and social research*. Berkshire: Open.
- Gosa, C. M. C. (2004). Investigating washback: A case study using student diaries. Unpublished doctoral dissertation, Lancaster University.
- Gradwell, J. M. (2006). Teaching in spite of, rather than because of, the test: A case of ambitious history teaching in New York State. In S. G. Grant (Ed.), *Measuring history: Cases of state-level testing across the United States* (pp. 157-176). Greenwich, CT: Information Age Publishing.
- Grant, C. A. (2004). Oppression, privilege, and high-stakes testing. *Multicultural Perspectives*, 6(1), 3-11.
- Green, A. (2006). Watching for washback: Observing the influence of the International English Language Testing System academic writing test in the classroom. *Language Assessment Quarterly* 3, 333–368.
- Green, A. (2007). Washback to learning outcomes: a comparative study of IELTS preparation and university pre-sessional language courses. *Assessment in Education: Principles, Policy & Practice*, 14(1), 75-97.
- Green, A. (2013). Washback in language assessment. *International Journal of English Studies*, 13(2), 39-51.
- Green, A. 2014. *Exploring language assessment and testing: Language in Action*. London: Routledge.

- Gregory, K., & Clarke, M. (2003). High-stakes assessment in England and Singapore. *Theory into practice*, 42(1), 66-74.
- Greive, E. L. (2012). *Comparing Alignment of a State Test and District Formative Assessments with State Content Standards using Three Methods* (Doctoral dissertation, The University of North Carolina at Chapel Hill).
- Griffin, M. L. (2003). Using critical incidents to promote and assess reflective thinking in pre-service teachers. *Reflective Practice*, 4(2), 207-220.
- Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Newbury Park, CA: Sage.
- Guerrero, M. D. (2000). The unified validity of the Four Skills Exam: applying Messick's framework. *Language testing*, 17(4), 397-421
- Guskey, T. R. (2007). Closing achievement gaps: Revisiting Benjamin S. Bloom's "Learning for mastery". *Journal of Advance Academics*, 19(1), 8- 31.
- Haertel, E., & Herman, J. (2005). A Historical Perspective on Validity: Arguments for Accountability Testing. CSE Report 654. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST)*.
- Hakim, C. (2000). *Research Design: Successful Designs for Social and Economic Research* (2nd edn). London: Routledge.
- Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed). Lawrence Erlbaum Associates, Mahwah, N.J
- Haladyna, T. M., & Downing, S. M. (2011). Twelve steps for effective test development. In *Handbook of test development* (pp. 17-40). Routledge.

- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education, 15*(3), 309-333.
- Haladyna, T. M., Nolen, S. B., & Haas, N. S. (1991). Raising standardized achievement test scores and the origins of test score pollution. *Educational Researcher, 20*(5), 2-7.
- Hamp-Lyons, L. (1997). Washback, impact and validity: Ethical concerns. *Language testing, 14*(3), 295-303.
- Hamp-Lyons, L. (1998). Ethical test preparation practice: The case of the TOEFL. *TESOL Quarterly, 32*(2), 329-337.
- Hamp-Lyons, L. (2002). The Scope of Writing Assessment. *Assessing Writing, 8*(1), 5-16.
- Han, B., Dai, M. & Yang L. (2004). Analyzing the problems of the College English Test based on a survey. *Foreign Language and Their Teaching, 179*(2), 17-23.
- Hansche, L. N. (1998). Meeting the requirements of Title I: Handbook for the development of performance standards. *Washington, DC: US Department of Education, 26*.
- Hargreaves, A. (1989). *Curriculum and assessment reform*. Toronto: OISE Press.
- Hargreaves, A. (1993). Individualism and individuality: Reinterpreting the teacher culture. In L. W. Little & M.W. McLaughlin (Eds.), *Teachers' work: Individuals, colleagues, and contexts* (pp. 51-76). New York: Teachers College Press.
- Hargreaves, A. (1994). *Changing teachers, changing times: Teachers' work and culture in the postmodern age*. Teachers College Press.

- Hargreaves, A., Earl, L., & Schmidt, M. (2002). Perspectives on alternative assessment reform. *American Educational Research Journal*, 39(1), 69-95.
- Harmer, J. (1991). *The practice of English language teaching*. London: Longman.
- Hashim, S. (1997). Review of Teaching English in Libya-Textbooks Used in Preparatory and Secondary Levels. In ESP in the Arab World: Reality Check and Prospects Proceedings of the XVIIth MATE Annual Conference Erfoud.
- Hashweh, M. (2003). Teacher accommodative change. *Teaching and Teacher Education*, 19(4), 421-434.
- Hawkey, R. A. (2006) *Impact theory and practice: studies of the IELTS test and Progetto Lingue 2000*. *Studies in Language Testing* 24 (Cambridge, Cambridge ESOL/ Cambridge University Press).
- Hayes, B., & Read, J. (2004). IELTS test preparation in New Zealand: Preparing students for the IELTS academic module. In L. Cheng, & Y. Watanabe with A. Curtis (Eds.), *Washback in language testing: Research contexts and methods*, (pp. 97-111). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.
- Heaton, J. B. (1990). *Classroom Testing*. Harlow, Longman.
- Heaton, J. B. (1990). *Longman Handbooks for Language Teachers*. New York: Longman Inc.
- Herman, J. L. (2004). The effects of testing on instruction. In S. H. Fuhrman & R. F. Elmore (Eds.), *Redesigning accountability systems for education* (pp. 141–166). New York: Teachers College Press.

- Herman, J. L., Webb, N. M., & Zuniga, S. A. (2007). Measurement issues in the alignment of standards and assessments. *Applied Measurement in Education*, 20(1), 101-126.
- Hess, K., Carlock, D., Jones, B., & Walkup, J. (2009). What exactly do “fewer, clearer, and higher standards” really look like in the classroom? Using a cognitive rigor matrix to
- Hmelo, C. E., & Ferrari, M. (1997). The problem-based learning tutorial: Cultivating higher order thinking skills. *Journal for the Education of the Gifted*, 20(4), 401-422.
- Hoadjli, A. C. (2015). *The Washback Effect of an Alternative Testing Model on Teaching and Learning: An Exploratory Study on EFL Secondary Classes in Biskra* (Doctoral dissertation, Université Mohamed Khider-Biskra).
- Holden, R.B. (2010). Face validity. In I.B. Weiner & W.E. Craighead (Eds.), *The corsini encyclopedia of psychology* (pp. 637–638). Hoboken, NJ: Wiley.
- Hoover, W. & Gough, P. (2013). The Reading Acquisition Framework - An Overview. Retrieved on April 15<sup>th</sup> 2018 from <http://www.sedl.org/reading/framework/framework.pdf>
- Hubley, N. J. (2012) Assessing reading. In: Coombe, C., Davidson, P., O’Sullivan, B. and Stoyhoff, S. (eds.) *The Cambridge guide to second language assessment* (pp.211-217). Cambridge: Cambridge University Press.
- Huebert, J. P., & Hauser, R. M. (Eds.). (1998). *High-stakes testing for tracking, promotion, and graduation*. Washington, DC: National Academy Press.

- Hughes, A. (1988). Introducing a needs-based test of English language proficiency into an English medium university in Turkey. In A. Hughes (Ed.), *Testing English for university study (ELT Documents #127)* (pp. 134-146). London: Modern English Publications in association with the British Council.
- Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.
- Hughes, A. (2003) *Testing for language teachers*, 2nd ed. Cambridge: Cambridge University Press.
- Impara, J. C. (2001, April). Alignment: One element of an assessment's instructional utility. In *annual meeting of the National Council on Measurement in Education, Seattle, WA* (pp. 1-13).
- Ingulsrud, John E. (1994). An entrance test to Japanese universities: social and historical context. In C. Hill & K. Parry (Eds.), *From testing to assessment: English as an international language* (pp. 61-81). New York: Longman.
- Izumi, S. (2003). Comprehension and production processes in second language learning: In search of the psycholinguistic rationale of the output hypothesis. *Applied Linguistics*, 24(2), 168-196.
- Jackson, P. W. (1992). Conceptions of curriculum and curriculum specialists. In P.W. Jackson (Ed) (1996). *Handbook of research on curriculum: A project of the American Educational Research Association*. (pp. 3-40). Macmillan Library Reference USA; Simon & Schuster Macmillan.

- Jacob, B. & Levitt, S. (2003). Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating. *Quarterly Journal of Economics*, 118 (3), 843-877.
- Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P., & Taylor, C. (2000). *TOEFL 2000 Framework*. Princeton, NJ: Educational Testing Service.
- Jerald, C.D. (2006). The hidden costs of curriculum narrowing. Washington, DC: Learning Point Associates Issue Brief. The Center for Comprehensive School Reform and Improvement. Retrieved April 31, 2010, from <http://files.eric.ed.gov/fulltext/ED494088.pdf>
- Jin, Y. (2000). The Washback Effect of the CET-SET on Teaching and Learning. *Foreign Language World*, 80 (4), 56-61.
- Jin, Z. (1990). *The imperial examination system and Chinese culture*. Shanghai: Shanghai People's Publishing House.
- Johnson, D. W., Johnson, R. T., & Stanne, M. E. (2000). *Cooperative learning methods: A meta-analysis*. Minneapolis, MN: University of Minnesota Press.
- Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational researcher*, 33(7), 14-26.
- Jongsma, K. S. (1993). Standards: Powerful Tools or Unnecessary Provocations? (Research to Practice). *Reading Teacher*, 46(4), 340-41.
- Kane, M. (1992). An argument-based approach to validity. *Psychological bulletin*, 112(3), 527-535.
- Kane, M. (2006). Content-related validity evidence in test development. *Handbook of test development*, 131-153. Lawrence Erlbaum Associates Publishers.

- Kaplan, B., & Maxwell, J. A. (2005). Qualitative research methods for evaluating computer information systems. In *Evaluating the organizational impact of healthcare information systems* (pp. 30-55). Springer, New York, NY.
- Kellaghan, T., & Greaney, V. (1992). *Using Examinations To Improve Education: A Study in Fourteen African Countries. World Bank Technical Paper Number 165. Africa Technical Department Series*. Distribution Unit, Office of the Publisher, Department F, The World Bank, 1818 H Street, NW, Washington, DC 20433 (free).
- Kentli, F. D. (2009). Comparison of hidden curriculum theories. *European Journal of Educational Studies, 1*(2), 83-88.
- Khaniya, T.R. (1990). The washback effect of a textbook-based test. *Edinburgh Working Papers in Applied Linguistics 1*. Edinburgh: University of Edinburgh.
- Kincheloe, J. L. (2001). Describing the bricolage: Conceptualizing a new rigor in qualitative research. *Qualitative inquiry, 7*(6), 679-692.
- Kincheloe, J. L., McLaren, P., & Steinberg, S. R. (2011). Critical pedagogy and qualitative research: Moving to the Bricolage. In N. K., Denzin & Y. S. Lincoln (4ed.). *The SAGE handbook of qualitative research. The Sage handbook of qualitative research*, (pp. 163-177), Sage.
- King, N. & Horrocks, C. (2010). *Interviews in qualitative research*. London: Sage.
- King, N. (1998) Template analysis. In G. Symon & C. Cassell (Eds.), *Qualitative methods and analysis in organizational research: A practical guide* (pp. 118-134). London: Sage.

- Kiss-Gulyás, J. (2001). Experiencing the examination design, content, materials and procedures. *English Language Education in Hungary, Part III: Training teachers for new examinations*, Együd, JG, Gál, IA and Author, P.(eds.). Budapest: The British Council, 40-58.
- Kliebard, H. M. (1989). Problems of Definition in Curriculum. *Journal of Curriculum and Supervision*, 5(1), 1-5.
- Koretz, D. M., Mitchell, K. J., Barron, S., & Keith, S. (1996). Perceived effects of the Maryland State Assessment Program (CSE Tech. Rep. No. 406). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
- Koretz, D., & Hamilton, L. S. (2006). Testing for accountability in K-12. In R. L. Brennan (Ed.) *Educational Measurement* (PP. 531 – 578). Westport: American Council on Education & Praegar.
- Koretz, D., Mitchell, K., Barron, S., & Keith, S. (1996). Final report: Perceived effects of the Maryland school performance assessment program. *Los Angeles, CA: CRESST*.
- Kvale, S. (2007). *Doing interviews*. Los Angeles: SAGE Publications Inc.
- La Marca, P. M. (2001). Alignment of standards and assessments as an accountability criterion. ERIC Development Team. (ERIC Document Reproduction Service No. ED458288)
- La Marca, P. M., Redfield, D., & Winter, P. C. (2000). *State Standards and State Assessment Systems: A Guide to Alignment*. Series on Standards and Assessments.

- La Marca, P. M., Redfield, D., Winter, P. C., & Despriet, L. (2000). State standards and state assessment systems: A guide to alignment. Series on standards and assessments. Washington, DC: Council of Chief State School Officers.
- La Marca, P. M., Redfield, D., Winter, P. C., Bailey, A., & Hansche, D. (2000). *State standards and state assessment systems: A guide to alignment*. Washington: Council of Chief State School Officers.
- Lam, H. P. (1993). Washback - Can it be qualified? A study on the impact of English examinations in Hong Kong. Unpublished master's thesis, University of Leeds, UK.
- Lam, H. P. (1994). Methodology washback - an insider's view. In D. Nunan, R. Berry & V. Berry (Eds.), *Bringing about change in language education: Proceedings of the international language in education conference 1994* (pp. 83-102). Hong Kong: University of Hong Kong.
- Lan, Y. F., Hung, C. L., & Hsu, H. J. (2011). Effects of guided writing strategies on students' writing attitudes based on media richness theory. *Turkish Online Journal of Educational Technology-TOJET*, 10(4), 148-164.
- Lane, S. (2004). Validity of high-stakes assessment: are students engaged in complex thinking? *Educational Measurement: Issues and Practice*, 23(3), 6-14.
- Larsen-Freeman, D. (2000). *Techniques and principles in language teaching* (2<sup>nd</sup> ed.). Oxford: Oxford University Press.
- Latham, H. (1877). *On the action of examinations considered as a means of selection*.

- Lee, J. F. (2014). A hidden curriculum in Japanese EFL textbooks: Gender representation. *Linguistics and Education, 27*, 39-53.
- Lee, L. (1994). L2 writing: Using pictures as a guided writing environment. Paper presented at the Rocky Mountain Modern Language Association. ERIC Document 386951.
- Leung, C. Y., & Andrews, S. (2012). The mediating role of textbooks in high-stakes assessment reform. *ELT journal, 66*(3), 356-365.
- Lewis, A. (2000). *High-stakes testing: Trends and issues*. Aurora, CO: Mid-Continent Research for Education and Learning, Policy Brief.
- Li, J. (2006). *Introducing audit trails to the world of language testing*. M. A. thesis, University of Illinois.
- Li, S., & Sireci, S. G. (2004). Evaluating the fit between test content, instruction, and curriculum frameworks: A review of methods for evaluating test alignment (Center for Educational Assessment Research Report No. 558). Amherst, MA: University of Massachusetts, Center for Educational Assessment.
- Li, X. (1990). How powerful can a language test be? The MET in China. *Journal of Multilingual and Multicultural Development, 11*(5), 393-404.
- Lingard, B., Sellar, S., & Lewis, S. (2017). Accountabilities in schools and school systems. In G. Noblit (Ed.), *Oxford Research Encyclopaedia of Education* (pp. 1-28). New York: Oxford University Press. <http://dx.doi.org/10.1093/acrefore/9780190264093.013.74>
- Linn, R. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4-16.

- Linn, R. L. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher*, 23(9), 4-14.
- Linn, R. L. (2005). Issues in the design of accountability systems. In E. H. Haertel & J. L. Herman (Eds.), *Uses and misuses of data for educational accountability and improvement*. The one hundred and fourth yearbook of the National Society for the Study of Education (part 2, pp. 78–98). Chicago: National Society for the Study of Education.
- Linn, R. L. (2006). Issues in the design of accountability systems. In J. L. Herman, & E. H. Haertel (Eds.), *Uses and misuses of data for educational accountability and improvement* (pp. 78-98). Chicago: National Society for the Study of Education.
- Lipman, P. (2004). *High stakes education*. New York: Routledge Falmer.
- Lobascher, S. (2011). What are the potential impacts of high-stakes testing on literacy education in Australia? *Literacy learning: The middle years*, 19(2), 9.
- Looney, J. W. (2009). *Assessment and Innovation in Education* (No. 24). OECD Publishing.
- Luna, C., & Turner, C. L. (2001). The impact of the MCAS: Teachers talk about high-stakes testing. *The English Journal*, 91(1), 79-87.
- Luxia, Q. (2007). Is testing an efficient agent for pedagogical change? Examining the intended washback of the writing task in a high-stakes English test in China. *Assessment in Education*, 14(1), 51-74.
- Lynch, B. K., & Davidson, F. (1994). Criterion-Referenced Language Test Development: Linking Curricula, Teachers, and Tests. *TESOL Quarterly*, 28(4), 727-743.

- Mackenzie, N., & Knipe, S. (2006). Research dilemmas: Paradigms, methods and methodology. *Issues in educational research*, 16(2), 193-205.
- Madaus, G. F. (1988). The influence of testing on the curriculum. In L. N. Tanner (Ed.), *Critical issues in curriculum: Eighty-seventh year book of the national society for the study of education* (pp. 83-121). Chicago: University of Chicago Press.
- Madaus, G., Russell, M., & Higgins, J. (2009). The paradoxes of high-stakes testing: How they affect students, their parents, teachers, principals, schools, and society. Charlotte, NC: Information Age Publishing
- Maghur, A. (2010). Highly-skilled migration (Libya): legal aspects. *CARIM analytic and synthetic Notes* 2010/31. Highly-Skilled Migration Series. Retrieved from <http://cadmus.eui.eu/dspace/handle/1814/13685>
- Mahgoub, M. M. (1977). *An Investigation of Some Specific English Phonological and Grammatical Problems: Of Twelfth-grade Literary Day-school Libyan Students of English as a Foreign Language*. (Doctoral dissertation, University of Kansas).
- Mahmoudi, L. (2014). *The washback effect of Iranian National University entrance exam (inuee) on pre-university English teaching and learning* (Doctoral dissertation, University of Malaya).
- Marchant, G. J. (2004). What is at stake with high stakes testing? A discussion of issues and research (1). *The Ohio Journal of Science*, 104(2), 2-8.
- Markee, N. (1997). *Managing curricular innovation*. Cambridge: Cambridge University Press.

- Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research*, 79(4), 1332-1361.
- Mays, N., & Pope, C. (2000). Assessing quality in qualitative research. *BMJ*, 320, 50-52.
- McCallum, B., Gipps, C., McAlister, S., & Brown, M. (1995). National curriculum assessment: Emerging models of teacher assessment in the classroom. In H. Torrance (Ed.), *Evaluating Authentic Assessment* (pp. 88–104). Philadelphia: Open University Press.
- McCallum, B., Gipps, C., McAlister, S., & Brown, M. (1995). National curriculum assessment: Emerging models of teacher assessment in the classroom. In H. Torrance (Ed.), *Evaluating authentic assessment: Problems and possibilities in new approaches to assessment* (pp.88–104). Buckingham, England: Open University Press
- McKay, P. (2007). The standards movement and ELT for school-aged learners: Cross national perspectives. In J. Cummins & C. Davison (Eds.), *International handbook of English language teaching: Part one* (pp. 439–456). New York: Springer.
- McKay, P., Coppari, P., Cumming, A., Graves, K., Lopriore, L., & Short, D. (2001). Language standards: An international perspective, part 1. *TESOL Matters*, 11(2), 1, 4.
- McNamara, T. (2000). *Language Testing*. Oxford: Oxford University Press.
- McNamara, T. (2001). Language assessment as social practice: Challenges for research. *Language Testing*, 18(4), 333-349.
- McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA: Blackwell.

- McNeil, L. (2002). *Contradictions of school reform: Educational costs of standardized testing*. Routledge.
- Menges, R. J. (1997). Fostering faculty motivation to teach: Approaches to faculty development. *Teaching Well and Liking It: Motivation Faculty to Teach Effectively*. Baltimore and London: The John Hopkins University.
- Mercer, N. (1995). *The guided construction of knowledge: Talk amongst teachers and learners*. Clevedon: Multilingual Matters.
- Mercer, N. (2000). *Words and minds: How we use language to think together*. London: Routledge.
- Meredith, J. (1998). Building operations management theory through case and field research. *Journal of operations management*, 16(4), 441-454.
- Merriam, S. B. (1988). *Case study research in education: A qualitative approach*. San Francisco: Jossey Bass.
- Mertens, D. M. (2005). *Research and evaluation in education psychology: Integrating diversity with quantitative, qualitative, and mixed methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Messick, S. (1989). Validity. In Linn, R. L. (Ed.), *Educational measurement*, 3rd edn. (pp. 13–103). New York: American Council on Education and Macmillan.
- Messick, S. (1996). Validity and washback in language testing. *Language Testing* 13, 241- 256.
- Miles, M. B., Huberman, A. M., & Saldana, J. (2014). *Qualitative Data Analysis: A Methods Sourcebook*. Sage Publications Ltd (CA).
- Minarechová, M. (2012). Negative impacts of high-stakes testing. *Journal of Pedagogy/Pedagogický Casopis*, 3(1), 82-100.

- Mirzaei, A., Hashemian, M., & Tanbakooei, N. (2012). Do Different Stakeholders' Actions Transform or Perpetuate Deleterious High-Stakes Testing Impacts in Iran? In *First Conference on Language Learning and Teaching: an Interdisciplinary approach*.
- Mizutani, S. (2009). *The mechanism of washback on teaching and learning*. Unpublished Ph.D. thesis, The University of Auckland, Auckland.
- Mohamed, S. (1987). *The Communicative Approach in Language Teaching and its Implications for Syllabus Design in Libya*. Doctoral Dissertation. University of East Anglia, United States.
- Morgan, D. L. (1998). Practical strategies for combining qualitative and quantitative methods: Applications to health research. *Qualitative health research*, 8(3), 362-376.
- Morris, T., & Wood, S. (1991). Testing the survey method: continuity and change in British industrial relations. *Work, Employment & Society*, 5(2), 259-282.
- Morrow, K. (1986). The evaluation of tests of communicative performance. In M. Portal (Ed.), *Innovations in language testing* (pp. 1-13). London: NFER/Nelson.
- Morse, J. M. (1991). Approaches to qualitative-quantitative methodological triangulation. *Nursing research*, 40(2), 120-123.
- Mullen, A. (2009). *The impact of using a proficiency test as a placement tool: The case of Test of English for International Communication (TOEIC)* (Doctoral Dissertation University of Laval).

- Näsström, G. (2008). *Measurement of alignment between standards and assessment*. Umeå, Sweden: Umeå universitet. Available from <http://umu.divaportal.org/smash/get/diva2:142244/FULLTEXT01>
- Näsström, G., & Henricksson, W. (2008). Alignment of standards and assessment: A theoretical and empirical study of methods for alignment. *Electronic Journal of Research in Educational Psychology*, 6(3), 667–690.
- Nazari, A. (2005). Washback effects on TEFL: A case study from Iran. *IATEFL VOICES*, 185, 9.
- Nichols, S. L., & Berliner, D. C. (2005). The inevitable corruption of indicators and educators through high-stakes testing (No. EPSL-0503-101-EPRU). Tempe, AZ: Education Policy Studies Laboratory, Arizona State University. Retrieved September 27, 2005.
- Nikolov, M. (1999). Classroom Observation Project. In H.Fekete, E. Major and M. Nikolov (eds.) *English language education in Hungary: a baseline study* (pp. 221-246). Budapest: The British Council Hungary.
- Nunan, D. (2007). Standards-based approaches to the evaluation of ESL instruction. In J. Cummins & C. Davison (Eds.), *International handbook of English language teaching: Part one* (pp. 421–438). New York: Springer.
- Nuthall, G., & Alton-Lee, A. (1995). Assessing classroom learning: How students use their knowledge and experience to answer classroom achievement test questions in science and social studies. *American Educational Research Journal*, 32(1), 185-223.
- O'Dwyer, S. (2006) The English teacher as facilitator and authority. *TESL-EJ*, Retrieved from <http://www-writing.berkeley.edu/TESL-EJ/ej36/a2.html>

- Olson, L. (2003). Standards and tests: Keeping them aligned. *Research points: Essential information for education policy*, 1(1), 1-4.
- Omar, Y. Z. (2014). *Perceptions of selected Libyan English as foreign language teachers regarding teaching of English in Libya* (Doctoral dissertation, University of Missouri-Columbia).
- Onaiba, E. M., & Mustafa, A. (2014). *Investigating the washback effect of a revised EFL public examination on teachers' instructional practices, materials and curriculum* (Doctoral Dissertation, University of Leicester).
- Onaiba, E.A. (2006) *An Evaluation of the Final English Examinations of Year Nine of the Basic Education Stage schools in Misrata, Libya*, unpublished M.A. dissertation. Academy of Postgraduate studies, School of Languages, English Department, Tripoli-Libya.
- Onwuegbuzie, A. J., Johnson, R. B., & Collins, K. M. (2009). Call for mixed analysis: A philosophical framework for combining qualitative and quantitative approaches. *International Journal of Multiple Research Approaches*, 3(2), 114-139.
- Orafi, S. M. S., & Borg, S. (2009). Intentions and realities in implementing communicative curriculum reform. *System*, 37(2), 243-253.
- Orfield, G., & Wald, J. (2000). Testing, testing. *The Nation*, 270(22), 38-40.
- Otman, W., & Karlberg, E. (2007). *The Libyan economy: economic diversification and international repositioning*. Springer Berlin Heidelberg: New York.
- Pajares, M. (1992). Teachers' beliefs and educational research: Cleaning up a messy construct. *Review of Educational Research*, 62(3), 307-332.

- Pan, Y. C. (2010). Test impact: English certification exit requirements in Taiwan. *TEFLIN Journal: A Publication on the Teaching and Learning of English*, 20(2).
- Paris, S. G. (2000). Trojan horse in the schoolyard. *Issues in Education*, 6(1/2), 1-16.
- Paris, S. G., & McEvoy, A. P. (2000). Harmful and enduring effects of high-stakes testing. *Issues in Education*, 6(1/2), 145-160.
- Patton, M. Q. (2002). *Qualitative Evaluation and Research Methods*. Thousand Oaks, CA: Sage Publications, Inc.
- Pearson, I. (1988) Tests as levers for change. In D. Chamberlain and R.J. Baumgardner (eds) *ESP in the classroom: practice and evaluation* (pp. 98–107). Modern English Publications, in association with the British Council.
- Pennycook, A. (1990). Towards a critical applied linguistics for the 1990s. *Issues in Applied Linguistics*, 1(1), 8-28.
- Perrone, V. (1991). On standardized testing. *Childhood Education*, 67(3), 131-142.
- Petrie, H. G. (1987). Introduction to "Evaluation and Testing". *Educational Policy*, 1(2), 175-180.
- Phelps, R. P. (2006). Characteristics of an effective student testing system. *Educational Horizons*, 85(1), 19-29.
- Pinter, Annamaria. (2007). Some benefits of peer-peer interaction: 10-year-old children practising with a communication task. *Language Teaching Research*, 11(2), 189-207.
- Plano Clark, V. L., Anderson, N., Wertz, J. A., Zhou, Y., Schumacher, K., & Miaskowski, C. (2015). Conceptualizing longitudinal mixed methods designs: a methodological review of health sciences research. *Journal of Mixed Methods Research*, 9(4), 297-319.

- Polesel, J., Dulfer, N., & Turnbull, M. (2012). The experience of education: The impacts of high stakes testing on school students and their families. *Sydney, Australia: The Whitlam Institute.*
- Popham, J. (1987). The merits of measurement-driven instruction. *Phi Delta Kappa, May*, 679–682.
- Popham, W.J. (2003), *Test Better, Teach Better: The Instructional Role of Assessment.* Association for Supervision and Curriculum Development, Alexandria, VA.
- Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational researcher, 31*(7), 3-14.
- Porter, A.C., & Smithson, J.L. (2001). Are content standards being implemented in the classroom? A methodology and some tentative answers. In S. Fuhrman (Eds.), *From the Capitol to the classroom: Standards-based reform in the states. One Hundredth Yearbook of the National Society for the Study of Education.* (pp. 60–80). Chicago: University of Chicago Press.
- Porter, A. C., & Smithson, J. L. (2002, April). Alignment of assessments, standards, and instruction using curriculum indicator data. In *Annual Meeting of the American Educational Research Association, New Orleans, LA.*
- Postholm, M. B. (2012). Teachers' professional development: a theoretical review. *Educational research, 54*(4), 405-429.
- Powell R.A., Single H.M., Lloyd K.R. (1996) 'Focus groups in mental health research: enhancing the validity of user and provider questionnaires', *International Journal of Social Psychology* 42 (3): 193-206.

Prabhu, N. S. (1988). *Second language pedagogy*. Oxford: Oxford University Press.

Prentice-Hall.

Prodromou, L. (1995) The Backwash Effect: from testing to teaching, *ELT Journal*, 49, pp. 13-25.

Pychyl, T. (2011). *A focus on students learning styles and students motivation*. [PowerPoint slides for ALDS5204W:PSYC6104W [16567:16501] Seminar in University Teaching (SEM) Winter 2011]. Retrieved from Lecture notes online website: <http://lms.carleton.ca/webct/cobaltMainFrame.dowebct>

Qi, L. (2004). Has a high-stakes test produced the intended changes. *Washback in language testing: Research contexts and methods*, 171-190.

Qi, L. (2005). Stakeholders conflicting aims undermine the washback function of a high-stakes test. *Language Testing*, 22(2), 142-173.

Qian, D. D. (2014). School-based English language assessment as a high-stakes examination component in Hong Kong: insights of frontline assessors. *Assessment in Education: Principles, Policy & Practice*, 21(3), 251-270.

Read, J., & Hayes, B. (2003). The impact of the IELTS on preparation for academic study in New Zealand. *IELTS International English Language Testing System Research Reports*, 4, 153-206.

Rea-Dickins, P. (1997). So, why do we need relationships with stakeholders in language testing? A view from the UK. *Language Testing*, 14(3), 304-314

- Reay, D. & Wiliam, D. (1999). 'I'll be a nothing': structure, agency and the construction of identity through assessment. *British Educational Research Journal*, 25(3), 343-354.
- Rensick, L.B., & Rensick, D. P. (1990). Tests as standards of achievement in school. In the uses of standardized tests in American education (pp. 63-80). Princeton, NJ: Educational Testing Service.
- Rentner, D. S., Scott, C., Kober, N., Chudowsky, N., Chudowsky, V., Joftus, S., & Zabala, D. (2006). From the capital to the classroom: Year 4 of the No Child Left Behind Act. *Washington, DC: Center on Education Policy*.
- Resnick, L. B., & Resnick, D. P. (1989). Tests as standards of achievement in school. In *Proceedings of the 1989 ETS Invitational Conference: The uses of standardized tests in American education* (pp. 63-80). Princeton, NJ: Educational Testing Service.
- Resnick, L. B., Rothman, R., Slattery, J. B., & Vranek, J. L. (2004). Benchmarking and alignment of standards and testing. *Educational Assessment*, 9(1-2), 1-27.
- Richards, C. (1998). The primary school curriculum: changes, challenges, questions. In: C. Richards & P.H. Taylor. (Eds), *How Shall We School Our Children? Primary education and its future*. London: Falmer Press.
- Roach, A. T., Elliott, S. N., & Webb, N. L. (2005). Alignment of an alternate assessment with state academic standards evidence for the content validity of the Wisconsin alternate assessment. *The Journal of Special Education*, 38(4), 218-231.

- Roach, A. T., Niebling, B. C., & Kurz, A. (2008). Evaluating the alignment among curriculum, instruction, and assessments: Implications and applications for research and practice. *Psychology in the Schools*, 45(2), 158-176.
- Roberts, M. (2000). *An examination of the way a group of Korean language learners prepare for the Test of English as a Foreign Language (TOEFL)*. (Unpublished Masters' Dissertation, University of Toronto).
- Roberts, M. (2000). *An examination of the way a group of Korean language learners prepare for the Test of English as a Foreign Language (TOEFL)*. Unpublished Masters' dissertation. Department of Curriculum, Teaching and Learning, University of Toronto.
- Rogers, E. M. (1983). *Diffusion of innovations*. 3rd edition, New York: Free Press.
- Ross, S. J. (2008). Language testing in Asia: Evolution, innovation, and policy challenges. *Language Testing*, 25(1), 5-13.
- Rothman, R., Slattery, J. B., Vranek, J. L., & Resnick, L. B. (2002). Benchmarking and alignment of standards and testing. *National Center for Research on Evaluation, Standards, and Student Testing*. Retrieved June 13, 2017 from <https://www.achieve.org/files/TR566.pdf>
- Ruch, G. M. (1929). *The objective or new-type examination*. Chicago, IL: Scott-Foresman.
- Ruhio, D. M., Berg-Weger, M., Yiehh, S. S., Lee, E. S., & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study. *Social Work Research*, 27, 2.
- Saif, S. (2006). Aiming for positive washback: A case study of international teaching assistants. *Language Testing*, 23(1), 1-34.

Sandelowski, M. (1993). Theory unmasked: The uses and guises of theory in qualitative research. *Research in Nursing & Health, 16*(3), 213-218.

Sandelowski, M. (1995). Qualitative analysis: What it is and how to begin. *Research in Nursing & Health, 18*(4), 371-375.

Saunders, M. N., Saunders, M., Lewis, P., & Thornhill, A. (2011). *Research Methods for Business Students, 5/e*. Pearson Education India.

Saville, N. (2009). *Developing a model for investigating the impact of language assessment within educational contexts by a public examination provider*. (Unpublished Doctoral Dissertation, University of Bedfordshire).

Saville, N., & Hawkey, R. (2004). The IELTS impact study: Investigating washback on teaching materials. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 73-96). New Jersey: Lawrence Erlbaum Associates.

Scaramucci, M. V. (2002). Entrance examinations and TEFL in Brazil: a case study. *Revista Brasileira de Linguística Aplicada, 2*(1), 1-13.

Schneider, A. L., & Ingram, H. M. (1997). *Policy design for democracy*. University Press of Kansas.

Scott, C. (2005). *Washback in the UK primary context with EAL learners: Exploratory case studies*. (Unpublished Doctoral Dissertation, University of Bristol).

Scott, C. (2007). Stakeholder perceptions of test impact. *Assessment in Education, 14*(1), 27-49.

Senior, R. (2006). *The experience of language teaching*. Cambridge: Cambridge

- Shafritz, J. M., Koeppe, R. P., & Soper, E. W. (1988). Dictionary of education. *Facts on File, New York*.
- Shepard, L. A. (1991). Will national tests improve student learning? *Phi Delta Kappan*, 232-238.
- Shih, C. M. (2006). Perceptions of the General English Proficiency Test and its Washback: A Case Study at Two Taiwan Technological Institutes. (Unpublished Doctoral Dissertation. University of Toronto).
- Shih, C. M. (2007). A new washback model of students' learning. *The Canadian Modern Language Review*, 64(1), 135-162.
- Shihiba, S. E. S. (2011). An Investigation of Libyan EFL Teachers' Conceptions of the Communicative Learner-Centred Approach in Relation to Their Implementation of An English Language Curriculum Innovation in Secondary Schools (Doctoral Dissertation, Durham University).
- Shohamy, E. (1992). Beyond proficiency testing: A diagnostic feedback testing model for assessing foreign language learning. *The Modern Language Journal*, 76(4), 513-521.
- Shohamy, E. (1993). *The power of test: The impact of language testing on teaching and learning*. Washington, DC: National Foreign Language Center.
- Shohamy, E. (1997). Testing methods, testing consequences: Are they ethical? Are they fair? *Language Testing*, 14(3), 340-349.
- Shohamy, E. (1998). Critical language testing and beyond. *Studies in Educational Evaluation*, 24(4), 331-345.

- Shohamy, E. (2000). Fairness in language testing. In Kunnan, A. J. (Ed.), *Fairness and validation in language assessment* (pp. 15–19). Cambridge, UK: Cambridge University Press.
- Shohamy, E. (2001). Democratic assessment as an alternative. *Language Testing*, 17(4), 373-391
- Shohamy, E. (2007). Tests as power tools: Looking back, looking forward. *Language testing reconsidered*, 141-152.
- Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: Washback effect over time. *Language Testing*, 13, 298-317.
- Sireci, S. G. (1998a). Gathering and analyzing content validity data. *Educational Assessment*, 5(4), 299-321.
- Sireci, S. G. (1998b). The construct of content validity. *Social indicators research*, 45(1-3), 83-117.
- Skrtic, T. (Ed.). (1995). *Disability and democracy: Reconstructing [special] education for post-modernity*. New York: Teachers College Press.
- Sloane, F. C., & Kelly, A. E. (2003). Issues in high-stakes testing programs. *Theory into Practice*, 42(1), 12-17.
- Smith, M. L. (1991). Put to the test: The effects of external testing on teachers. *Educational Researcher*, 20(5), 8-11.
- Smith, M. L., & Rottenberg, C. (1991). Unintended consequences of external testing in elementary schools. *Educational Measurement: Issues and Practice*, 10(4), 7-11.

- Spada, N., & Froehlich, M. (1995). COLT: *Communicative orientation of language teaching observation scheme, coding conventions and applications*. Sydney, NSW: Macquarie University, National Center for English Language Teaching and Research.
- Spolsky, B. (1994). The examination-classroom backwash cycle: Some historical cases. Paper presented at the ILEC'94, University of Hong Kong.
- Spratt, M. (2005). Washback and the classroom: The implications for teaching and learning of studies of washback from exams. *Language Teaching Research*, 9(1), 5-29.
- Stake, R. E. (2010). *Qualitative research: Studying how things work*. New York: The Guildford Press.
- Stalmeijer, R. E., McNaughton, N., & Van Mook, W. N. (2014). Using focus groups in medical education research: AMEE Guide No. 91. *Medical Teacher*, 36(11), 923-939.
- Statman, S. (1988). Ask a clear question and get a clear answer: An enquiry into the question/answer and the sentence completion formats of multiple-choice items. *System*, 16, 367-376.
- Stecher, B. M. (2002). Consequences of large-scale, high stakes testing on school band classroom practice. L. S. Hamilton, B. M. Stecher & S. P. Klein (Eds.). *Making sense of test-based accountability in education* (pp. 79-100). Santa Monica: RAND Corporation.
- Stecher, B. M., Barron, S. L., Chun, T., & Ross, K. (2000). *The effects of the Washington state education reform on schools and classrooms* (No. RAND/DRU-2263-1). Rand Corp Santa Monica CA.

- Steinberg, S. (2011). Introduction to “Describing the bricolage”. In JL Kincheloe, k. hayes, S. Steinberg, and K. Tobin (Eds.), *Key works in critical pedagogy (Vol. 32)*. p. 177, Springer Science & Business Media.
- Stiggins, R. J. (1999). Assessment, student confidence, and school success. *The Phi Delta Kappan*, 81(3), 191-198.
- Sturman, P. (1996). Registration and placement: Learner response. In K. M. Bailey & D. Nunan (Eds.), *Voices from the language classroom: Qualitative research in second language education* (pp. 338- 355). Cambridge: Cambridge University Press.
- Suwaed, H. H. (2011). *Teachers' cognition and classroom teaching practice: an investigation of teaching English writing at the university level in Libya* (Doctoral Dissertation, University of Glasgow).
- Tan, H. M. (2009). *Changing the language of instruction of mathematics and science in Malaysia: The PPSMI policy and washback effect of bilingual high-stakes secondary school exit exams* (Doctoral Dissertation, McGill University).
- Tan, M. (2008). Bilingual high-stakes mathematics and science exams in Malaysia: Pedagogical and linguistic issues. Paper presented at the Language Testing and Research Colloquium, Hangzhou, China.
- Tan, M., & Turner, C. E. (2015). The impact of communication and collaboration between test developers and teachers on a high-stakes ESL exam: Aligning external assessment and classroom practices. *Language Assessment Quarterly*, 12(1), 29-49.

- Tang, X. Y. (2005). The washback study of language testing. *Foreign languages and Their Teaching* 7, 55-59.
- Tantani, A. S. N. (2012). *Significant Relationships between EFL Teachers' Practice and Knowledge in the Teaching of Grammar in Libyan Secondary Schools*. (Unpublished Doctoral Dissertation, University of Sunderland).
- Tashakkori, A. and Teddlie, C. (1998). *Mixed methodology: combining qualitative and quantitative approaches*. Thousand Oaks, CA: Sage.
- Tashakkori, A., & Teddlie, C. (Eds.). (2003). *Handbook of mixed methods in social & behavioral research*. Thousand Oaks, CA: Sage.
- Taylor, G., Shepard, L., Kinner, F., & Rosenthal, J. (2002). A survey of teachers' perspectives on high-stakes testing in Colorado: what gets taught, what gets lost? Center for Research on Evaluation, Diversity and Excellence, University of California, Santa Cruz and Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles. ERIC Document Reproduction Service No. ED475139.
- Taylor, P. C., Fraser, B. J., & Fisher, D. L. (1997). Monitoring constructivist classroom learning environments. *International Journal of Educational Research*, 27, 293-302.
- Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research*. Thousand Oaks, CA: SAGE.
- Teddlie, C., & Yu, F. (2007). Mixed methods sampling: A typology with examples. *Journal Of Mixed Methods Research*, 1(1), 77-100.

- Thronbury, S. (1999). *How to teach grammar*. England: Pearson Education Limited.
- Thorndike, E. (1921). Measurement in education. *Teachers College Record*, 22(5), 371–379.
- Tienken, C., & Wilson, M. (2001). Using state standards and tests to improve instruction. *Practical Assessment, Research & Evaluation*, 7(3). Retrieved April 14, 2015 from <http://PAREonline.net/getvn.asp?v=7&n=13>
- Tishman, S., Jay, E., & Perkins, D. N. (1993). Teaching thinking dispositions: From transmission to enculturation. *Theory into Practice*, 32(3), 147-153.
- Trim, J. (1998). European perspectives on modern language learning: Contributions to the Modern Languages Project of the Council of Europe. *Language Teaching*, 31(2), 206–217.
- Tsagari, D. (2011). Washback of a high-stakes English exam on teachers' perceptions and practices. *Selected Papers on Theoretical And Applied Linguistics*, 19, 431-445.
- Tsagari, K. (2006). *Investigating the washback effect of a high-stakes EFL exam in the Greek context: Participants' perceptions, material design and classroom applications*. (Doctoral Dissertation, University of Lancaster).
- Tudor, I. (2001). *The dynamics of the language classroom*. Cambridge: Cambridge University Press.
- Turner, C. E. (2014). Mixed methods research. In A.J. Kunnan (Ed.), *The companion to language assessment* (pp. 1-15). Hoboken, NJ: John Wiley & sons, Inc.
- Turner, C.E. (2001). The need for impact studies of L 2 performance testing and rating: Identifying areas of potential consequences at all levels of the testing cycle. In A. Brown et al. (Eds.),

*Experimenting with Uncertainty: Language Testing Essays In Honour Of Alan Davies*, pp.138-149. Cambridge, Cambridge University Press.

Turner, C.E. (2005). Professionalism and high-stakes tests: Teacher perspectives when dealing with educational change introduced through provincial exams. *TESL Canada Journal*, 23 (2), 54-76.

Turner, C.E. (2008). The specificity of the “research approach” in classroom studies: Probing the predictability of washback through teacher conceptual and instrumental evidence in Quebec high schools. Paper presented at the Language Testing and Research Colloquium 2008, Hangzhou.

Turner, C.E. (2009). Examining washback in second language education contexts: A high stakes provincial exam and the teacher factor in classroom practice in Quebec secondary schools. *International Journal of Pedagogies and Learning*, 5(1), 103-123.

Turner, J., & Paris, S. G. (1995). How literacy tasks influence children's motivation for literacy. *The Reading Teacher*, 48(8), 662-673.

Tyack, D., & Cuban, L. (1995). *Tinkering toward utopia: A century of public school reform*. Cambridge, MA: Harvard University Press.

Tyler, R. W. (1949). *Basic principles of curriculum and instruction*. Chicago, IL: University of Chicago.

UK Embassy (2005) The UK and Libya: Our Shared Tradition [online]. Retrieved from <http://www.britishembassy.gov.uk/servlet/Front?pagename=OpenMarket/Xcelerate/ShowPage&c=Page&cid=1064572031651>

- Urduan, T. C., & Paris, S. G. (1994). Teachers' perceptions of standardized achievement tests. *Educational Policy*, 8(2), 137-156.
- Valli, L., & Buese, D. (2007). The changing roles of teachers in an era of high-stakes accountability. *American Educational Research Journal*, 44(3), 519-558.
- Van der Walt, J. L., & Steyn Jr, H. S. (2008). The validation of language tests. *Stellenbosch papers in linguistics*, 38(1), 191-204.
- Van Hover, S. D. (2006). Teaching history in the old dominion: The impact of Virginia's accountability reform on seven secondary beginning history teachers. In S. G. Grant (Ed.), *Measuring History: Cases of State-Level Testing Across The United States* (pp. 195-219). Greenwich, CT: Information Age Publishing
- Vogler, K. E. (2003). An integrated curriculum using state standards in a high-stakes testing environment. *Middle School Journal*, 34(4), 10
- Volante, L. (2004). Teaching to the test: what every educator and policy-maker should know. *Canadian Journal of Educational Administration and Policy*, 35, 1-6.
- Vygotsky, L. S. (1986). *Thought and language* (Rev. ed.). Cambridge: MIT Press.
- Vygotsky, L.S. (1991). *Pedagogicheskay psikhologia* [Pedagogical psychology]. (N, Artemeva, Trans.). Moscow: Pedagogica.
- Wall, D. (1996). Introducing new tests into traditional systems: Insights from general education and from innovation theory. *Language Testing*, 13(3), 334-354.

- Wall, D. (1997). Impact and washback in language testing. In C. Clapham & D. Corson (Eds.), *Encyclopaedia of language and education: Vol. 7. Language testing and assessment* (pp. 291-302). Dordrecht: Kluwer Academic.
- Wall, D. (1999). The impact of high-stakes examinations on classroom teaching: a case study using insights from testing and innovation theory. (Unpublished Doctoral Dissertation, University of Lancaster).
- Wall, D. (2005). *Studies in Language Testing*: Cambridge, UK: Cambridge University Press.
- Wall, D. (2012). Washback. The Routledge handbook of language testing. In G., Fulcher, & F, Davidson, *The Routledge handbook of language testing* (pp. 79-92.). Routledge.
- Wall, D., & Alderson, J. C. (1993). Examining washback: the Sri Lankan impact study. *Language Testing*, 10(1), 41-69.
- Wall, D., & Horák, T. (2007). The impact of changes in the TOEFL examination on teaching and learning in Central and Eastern Europe: Phase 2, Coping with change. *ETS Research Report Series*, 2008(2), i-105.
- Wang, H. (2006). An Implementation of the English as a Foreign Language Curriculum Policies in the Chinese Tertiary Context. (Doctoral Dissertation, Queen's University).
- Wang, J. (2010). *A Study of the Role of the 'teacher Factor' in Washback*. (Unpublished Doctoral Dissertation. University of McGill).
- Watanabe, Y. (1996). Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research. *Language Testing*, 13 (3), 318-333.

- Watanabe, Y (1997) *The Washback Effects of the Japanese University Entrance Examinations of English-Classroom-Based Research*, (Unpublished Doctoral Dissertation, University of Lancaster).
- Watanabe, Y. (2001). Does the university entrance examination motivate learners? A case study of learner interviews. In A. Murakami (Ed.), *Trans-equator exchanges: A collection of academic papers in honour of Professor David Ingram* (pp. 100–110). Akita, Japan: Akita University.
- Watanabe, Y. (2004). Methodology in washback studies. In L. Cheng, & Y. Watanabe with A. Curtis (Eds.), *Washback in language testing: Research contexts and methods*, (pp. 19-36). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.
- Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (NISE Research Monograph No. 6). Madison: University of Wisconsin-Madison, National Institute for Science Education.
- Webb, N. L. (1999). Alignment of Science and Mathematics Standards and Assessments in Four States. *Research Monograph* No. 18.
- Webb, N. L. (2002). *An analysis of the alignment between mathematics standards and assessments for three states*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Webb, N. L., Horton, M., & O’Neal, S. (2002, April). *An analysis of the alignment between language arts standards and assessments in four states*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

- Webb, N. M., Herman, J. L., & Webb, N. L. (2007). Alignment of mathematics state-level standards and assessments: The role of reviewer agreement. *Educational Measurement: Issues and Practice*, 26(2), 17-29.
- Wedell, M. (1992). Pre/in service training of ELT teacher trainers: Planning the regional MATEFL project. *Perspectives on second language teacher education*, 337-350.
- Weir, C. J. (2005). *Language testing and validation*. UK: Macmillan.
- West, A. (2010). High stakes testing, accountability, incentives and consequences in English schools. *Policy & Politics*, 38(1), 23-39.
- West, M. (2007). Testing, learning, and teaching: The effects of test-based accountability on student achievement and instructional time in core academic subjects. In C. F. Finn & D. Ravitch (Eds.), *Beyond basics: Achieving a liberal education for all children* (pp. 46–62). Washington, DC: Thomas B. Fordham Institute
- White, C.J. (1989). Negotiating communicative language learning in a traditional setting. *ELT Journal* 43, 213-220.
- Whitehead, A. J. (1989). Creating a living educational theory from questions of the kind, “How do I improve my practice? *Cambridge Journal of Education*, 19(1), 41–51.
- Wigfield, A., & Eccles, J. S. (1989). Test anxiety in elementary and secondary school students. *Educational Psychologist*, 24(2), 159-183.
- Wiliam, D. (1996). National curriculum assessments and programmes of study: validity and impact. *British Educational Research Journal*, 22(1), 129-141.

- Wiliam, D. (2001). An overview of the relationship between assessment and the curriculum. *Curriculum And Assessment*, 165-181.
- Williamson, P., Bondy, E., Langley, L., & Mayne, D. (2005). Meeting the challenge of high-stakes testing while remaining child-centered: The representations of two urban teachers. *Childhood Education*, 81(4), 190-195.
- Windh, C., & Gingell, J. (1999). *Key concepts in the philosophy of education*. London: Routledge.
- Winfield, L. F. (1993). Investigating test content and curriculum content overlap to assess opportunity to learn. *Journal of Negro Education*, 62, 288–310.
- Wolf, S. A., Wolf, K. P., & Carpenter, M. (2002). Teaching true and to the test in writing. *Language Arts*, 79(3), 229.
- Woods, D. (1996). *Teacher Cognition in Language Teaching*. Cambridge: Cambridge University Press.
- Xi, X. (2010). How do we go about investigating test fairness?. *Language Testing*, 27(2), 147-170.
- Xie, Q., & Andrews, S. (2013). Do test design and uses influence test preparation? Testing a model of washback with Structural Equation Modeling. *Language Testing*, 30(1), 49-70.
- Yin, R. K. (2009). *Case study research: Design and methods*, 4th. *Thousand Oaks*.
- Yu, Y. (2005). *Washback Effects of the Spoken English Test on Test Book Content*. Unpublished MA, Queen's University.

- Yu, Y. (2010). *The washback effects of school-based assessment on teaching and learning: A case study*. (Doctoral dissertation). Retrieved from <http://hub.hku.hk/bitstream/10722/65273/1/FullText.pdf?accept=1>
- Zafeiriadou, N. (2009). Drama in language teaching: A challenge for creative development. *Issues*, 23, 4-9
- Zarza, N., & Abedalazeez, N. (2013). Simple versus complex stems in multiple-choice tests and their effects on students' performance. *International Journal of Applied Linguistics & English Literature*. 3(2), 237-241.
- Zhou, Y. (2004). Comparability study of two National EFL Tests (CET-6 and TEM-4) in China. *The Journal of Asia TEFL*, 1(1), 75–100.

## Appendices

### Appendix A: Former Version of BECEE which is similar to the former SECEE (From Onaibia, 2014, pp 297-300)

#### References

---

#### Appendices

#### Appendix 1 A sample of the Old BECE in English

*Final Exam for Preparatory Schools*

*English Language Examination*

*First session, 2007*

*“Make sure that the exam includes 11 questions, and the number of pages is ‘5’” (Translated)*

**Time: 3 Hours**

**Total Marks: (60)**

---

**Q.1.: Read and answer the questions below:**

Libya is in north Africa. Its total area is 1.759.540 square kilometres and its population is about 5.5 million. Arabic is the official language in Libya, and some people also speak English or Italian. The country lies on the coast of the Mediterranean sea. This is also where the country's capital, Tripoli, is. Its neighbouring countries are Tunisia and Algeria in the west. Algeria is the second largest country in Africa. Its total area covers 2.382.740 square kilometres. There are about 28 million people in the country. The official language is Arabic and many people also speak French. The country's capital is Algiers.

Tunisia is a small country in north Africa, with a total area of 163.610 square kilometres and a population of about 9 million. Tunisia is further north than any other country in Africa. The country's capital is Tunis. The official language is Arabic. Many people speak French as a second language.

*Answer these questions:*

- a- What other language do some people in Libya speak?  
.....
- b- Which is smaller, Libya or Tunisia?  
.....
- c- Which country has a bigger population, Tunisia or Algeria?  
.....
- d- What are Libya's neighbouring countries in the west?  
.....
- e- Where is Tripoli?  
.....
- f- Which country is the furthest north?  
.....
- g- Is Algeria the biggest country in Africa?  
.....

(7 x 1 = 10 ½ marks)

## References

**Q.2.: Fill in the space with these words:**

|      |      |      |          |      |           |
|------|------|------|----------|------|-----------|
| walk | help | hurt | climbing | hour | mountains |
|------|------|------|----------|------|-----------|

Ali and Nuri went on a holiday to the ..... because they like ..... They climbed for about ..... Then Ali fell and ..... His leg. He couldn't ....., so Nuri went for .....

(6 x ½ = 3 marks)

**Q.3.: Choose the best answer:**

- a- There isn't (many – much) traffic in villages.
- b- There is (much – plenty) of entertainment.
- c- Do (many – much) people live in Tripoli.
- d- The girl (which – who) came is my sister.
- e- (Which – Who) cleaned the blackboard?
- f- This is the car (which – who) I want to buy.

(6 x ¼ = 3 marks)

**Q.4.: Complete with correct words and right spelling:**

- a- ..... come before Sunday.
- b- ..... comes after Tuesday.
- c- Today is Friday, yesterday was .....
- d- ..... is the first month of the year.
- e- June, July and ..... are summer.
- f- ..... comes before Monday.

(6 x 1 = 6 marks)

**Q.5.: Join these sentences as it is shown in the brackets.**

- a- I like tea. I don't like coffee. (but)  
.....
- b- I was five. I started school. (when)  
.....
- c- you can come here. You can stay at home. (or)  
.....
- d- he was happy. He told everybody. (so ..... that)  
.....

(4 x 1½ = 6 marks)

**Q.6.: Write in words:**

- a- 21<sup>st</sup> = .....
- b- 90<sup>th</sup> = .....
- c- 100<sup>th</sup> = .....
- d- 1/6/98 = .....
- e- 13/11/1735 = .....
- f- f-22/2/1542 = .....

(6 x 1 = 6marks)

*References*

---

**Q.7.: Put ( ) or ( ) in front of each sentence, then correct the wrong ones:**

- a- I work on a farm. I look after sick people. ( )
- b- A fall can be very dangerous for old people. ( )
- c- We use (be + going to + infinitive) to talk about intentions. ( )
- d- People around the world eat the same kind of food. ( )
- e- The first explorers we know about were from America. ( )
- f- Sudan is the largest country in Africa. ( )
- g- Naguib Mahfouz was born in Tokyo. ( )
- h- Mahfouz began writing at the age of seventeen. ( )
- i- If you put a cork in water it sinks. ( )
- j- If you water plants they die. ( )

.....

.....

.....

.....

.....

(10 x ½ = 5 marks)

---

**Q.8.: write the vowels to complete the words.**

- a- H... nggl...d...ng
- b- Sc...b...d...y...ng
- c- p... .. m
- d- p... ..no
- e- sp ...gh...tt...
- f- ... ..dh(6 x 1 = 6 marks)

**Q.9.: Put these words in the right space**

|           |            |             |       |         |
|-----------|------------|-------------|-------|---------|
| conductor | electrical | electricity | atoms | equator |
|-----------|------------|-------------|-------|---------|

- a- Salt water is good ..... of .....
- b- Lightning is a huge ..... spark in the sky.
- c- Tropical rainforests grow near the .....
- d- Everything around us is made up of .....

(5 x ½ = 2 ½ marks)

---

**Q.10.: Rewrite the sentences using "if"**

- a- Stay at home when there us a bad storm.  
If .....
- b- Don't touch anything electrical when you have wet hands.  
If .....
- c- Take plenty of water when you go camping in the desert.

*References*

---

- If .....
- a- Don't drive when the weather is very bad.  
.....
- b- Study hard the week before when you have a test.  
If .....
- c- Ask your father for help when you can't do your homework.  
If .....
- (6 x 1 = 6 marks)
- 

**Q.11: Write 6 sentences about your lifestyle.**

These questions may help you.

- a- What kind of food do you eat?  
b- Do you ever miss meals?  
c- Do eat between meals?  
d- Do eat too much?  
e- How often do take exercise?  
f- What kind of exercise do you take?

.....

.....

.....

.....

.....

(6 sentences, one mark for each)

**GOOD LUCK**

**Appendix B: A Sample Test of the rSECEE**

## The New Secondary Education Certificate Examination of English

True and false questions:

Q1. Bronze was first made in the middle east.

- C. True
- D. False

Q2. When we review a story, we use the present tense.

- A. True
- B. False

Q3. “Ashamed of “is a positive word.

- A. True
- B. False

Q4. Early learning should continue to be rewarded from time to time.

- C. True
- D. False

Q5. A person who watches a sporting team is a spectator.

- A. True
- B. False

Q6. In reported speech “last night” must be changed to “that night”.

- A. True
- B. False

Q7. “May” and “Might” have a different meaning.

- A. True
- B. False

Q8. “Conditioning” is a learning process.

- A. True
- B. False

Q9. He said he had done it, yesterday.

- A. True
- B. False

Q10. American people use the present perfect tense less often.

- A. True
- B. False

Q11. The Nazca lines are parallel and intersect.

- A. True
- B. False

Q12. “Apparently” is used when we are sure about something.

- A. True
- B. False

Q13. They would have won if they had worked hard.

- A. True
- B. False

Q14. The winter in Oman is like a European winter.

- A. True
- B. False

Q15. Copper, gold, and silver are hard metals.

- A. True
- B. False

Q16. I'll be taking my final exam next Sunday.

- A. True
- B. False

Q17. We expected him coming soon.

- A. True
- B. False

Q18. The World Cup is a knockout competition.

- A. True
- B. False

Q19. Smallpox was a thing of the past.

- A. True
- B. False

Q20. The lake is enough warm to swim in.

- A. True
- B. False

Q21. Temperatures can create dust storms.

- A. True
- B. False

Q22. The nutrients are used to provide us with energy.

- A. True
- B. False

Q23. The word “socialize” is a verb.

- A. True
- B. False

Q24. My uncle is kind for me.

- A. True
- B. False

Q25. Sandy soil is better for plants than fertile soil.

- A. True
- B. False

Multiple choice questions

Q26. A .....is an association of sporting clubs.

- A. Team
- B. Club
- C. Group
- D. League

Q27. Coal is a .....resource.

- A. Living
- B. Non living
- C. Renewable
- D. Non renewable

Q28. The World Cup .....is a huge global business.

- A. Is become
- B. Was become
- C. Has become
- D. Have become

Q29. I'm never late .....school.

- A. in
- B. on
- C. at
- D. for

Q30. Psychologists use .....to compare people's personalities.

- A. equipment
- B. operations
- C. tests
- D. tools

Q31. The ancient Egyptians used .....to build the pyramids.

- A. marble
- B. concrete
- C. stone
- D. grain

Q32. The winter in Oman is like a European.....

- A. Winter
- B. Spring
- C. Summer
- D. Autumn

Q33. I wished we .....another way.

- A. go

- B. went
- C. have gone
- D. had gone

Q34. ....can be found in the meat and fish to build our muscle.

- A. Proteins
- B. Carbohydrates
- C. Fats
- D. Vitamins

Q35. The Red Crscent was set up .....years ago.

- A. few
- B. a few
- C. many
- D. much

Q36. There are .....types of Arabic language.

- A. two
- B. three
- C. four
- D. five

Q37. ....dinosauars were meat eater.

Some

- A. Few
- B. Little
- C. All

Q38. ....billion people in the world speak English fluently)

- A. One
- B. Two
- C. Three
- D. Four

Q39. “Can’t” is often used in passive sentences in the .....

- A. past
- B. present
- C. future
- D. near future.

Q40. The Nazca lines can be seen only .....a plane.

- A. in
- B. on
- C. at
- D. from

Q41. “David Copperfield” is set mainly in ..... .

- A. Paris
- B. Rome
- C. Berlin
- D. London

Q42. ....stories are written in the past tense. .

- A. Few
- B. Some
- C. Most
- D. All

Q43. Is she ready.....her.

- A. for
- B. with
- C. about
- D. to

Q44. Our bodies tell us ..... we need.

- E. what
- F. who
- G. whom

H. where

Q45. The.....was the largest amphitheatre in the Rome world.

- A. Colosseum
- B. Taj Mahal
- C. Pyramid
- D. Castle

Q46. .... Is made of rocks and plant materials.

- A. Coal
- B. Concrete
- C. Soil
- D. Bronze

Q47. “bored” is used to say .....people feel.

- A. which
- B. how
- C. what
- D. who

Q48. The Colosseum was.....by the Romans. .

- A. build
- B. built
- C. builded
- D. building

Q49. In reported speech “today” must be changed to .....

- A. that day
- B. the next day
- C. the previous day
- D. the following day

Q50. Rousseau outlined his theory with .....stages of development. three

- A. four
- B. five
- C. six

Match column A with its correspondence from column B

| A                         | B                      |
|---------------------------|------------------------|
| Q51. I was happy          | A. to do here          |
| Q52. There's nothing      | B. to start the lesson |
| Q.53 It's too heavy       | C. to trouble you      |
| Q54. I'm sorry            | D. to hear the news    |
| Q55. The teacher is ready | E. To work on it       |

|                   |                          |
|-------------------|--------------------------|
| Q56. Claim        | A. Looked carefully      |
| Q57. Surprisingly | B. Working together      |
| Q58.Loves ones    | C. Say something is true |
| Q59.Peered        | D. It seems strange      |
| Q60. Cooperation  | E. Family members        |

Exam ended

### Appendix C: Review on Previous and Current Washback Studies

Summary table of some previous and current washback studies

| Researchers                                | Context       | Exam   | Methodology employed & methods used   |
|--|---------------|--|---|
| Wall & Alderson (1993)                     | Sri Lanka     | O-level Examination in English   | Mixed methods <ul style="list-style-type: none"> <li>• Questionnaires</li> <li>• Interviews</li> <li>• Observations</li> <li>• Materials and test Analysis</li> </ul>       |
| Lam (1993)                                 | Hong Kong     | Revised use of English Exam 1989   | Quantitative <ul style="list-style-type: none"> <li>• Questionnaires to teachers</li> <li>• Analysis of textbooks</li> <li>• Analysis of test scripts and scores</li> </ul> |
| Andrews (1994)                             | Hong Kong     | Oral component of the revised use of English (RUE)                         | Quantitative <ul style="list-style-type: none"> <li>• Questionnaires</li> </ul>   |
| Andrews & Fullilove (1994)                 | Hong Kong     | RUE  | Qualitative: <ul style="list-style-type: none"> <li>• Interviews</li> </ul>   |
| Ingulsurd (1994)                           | Japan         | University Entrance test   | Qualitative: <ul style="list-style-type: none"> <li>• Observation</li> <li>• Interviews</li> </ul>  |
| Alderson & Hamp-Lyons (1996)               | United States | TOEFL  | Qualitative <ul style="list-style-type: none"> <li>• Interviews</li> <li>• Observations</li> <li>• Field notes</li> <li>• Audio recordings</li> </ul>                       |
| Shohamy, Dantsa – Schmidt & Freeman (1996) | Israel        | Arabic as a second language test<br><br>English as a foreign language test | Qualitative <ul style="list-style-type: none"> <li>• Questionnaires</li> <li>• Interviews</li> <li>• Analysis of inspectorate bulletins</li> </ul>                          |

|                                   |             |  |  |
|-----------------------------------|-------------|--|--|
| Watanabe (1996)                   | Japan       | University entrance exam   | Mixed Methods <ul style="list-style-type: none"> <li>• Analysis of textbooks and teaching material</li> <li>• Observations and interviews with two teachers</li> </ul> |
| Sturman (1996)                    | Japan       | Placement test   | Mixed methods: <ul style="list-style-type: none"> <li>• Questionnaire</li> <li>• Ethnographic</li> </ul>   |
| Cheng (1997, 1998)                | Hong Kong   | Revised Hong Kong Certificate of Education Examinations in English (HKCEE).  | Mixed Methods <ul style="list-style-type: none"> <li>• Classroom observations</li> <li>• Teacher and student questionnaires</li> <li>• Interviews</li> </ul>           |
| Cheng (1999)                      | Hong Kong   | HKCEE  | Qualitative <ul style="list-style-type: none"> <li>• Observations</li> <li>• Interviews</li> </ul>   |
| Hamp Lyons (1998)                 | Hong Kong   | TOFEL  | Qualitative <ul style="list-style-type: none"> <li>• Analysis of 5 TOEFL preparation textbooks</li> </ul>  |
| Nikolov (1999)                    | Hungary     | EFL Hungarian Examination  | Qualitative: <ul style="list-style-type: none"> <li>• Observations</li> <li>• Interviews</li> </ul>  |
| Lumely & Stoneman (2000)          | Hong Kong   | Graduating students' Language Proficiency Assessment                         | <ul style="list-style-type: none"> <li>• Interviews</li> <li>• Questionnaires</li> </ul>   |
| Andrews, Fullilove, & Wong (2002) | Hong Kong   | Hong Kong Advanced Supplementary (AS) 'Use of English' (UE) oral examination | Mixed Methods: <ul style="list-style-type: none"> <li>• Video-taping of student performance</li> <li>• Interviews</li> </ul>   |
| Read & Hayes (2003)               | New Zealand | IELTS  | Mixed methods: <ul style="list-style-type: none"> <li>• Questionnaires</li> <li>• Interviews</li> <li>• Observations</li> <li>• Pre-post tests</li> </ul>              |

|                    |                |  |  |
|--------------------|----------------|--|--|
| Ferman (2004)      | Israel         | EFL Oral Matriculation Test                    | Mixed methods: <ul style="list-style-type: none"> <li>• Structured questionnaires</li> <li>• Structured interviews</li> <li>• Open interviews</li> <li>• Document analysis</li> </ul>                                      |
| Qi (2004)          | China          | The National Matriculation English Test (NMET) | Mixed methods: <ul style="list-style-type: none"> <li>• Questionnaires</li> <li>• Interviews</li> </ul>  |
| Watanabe (2004)    | Japan          | Japanese University Entrance Examination       | Mixed methods: <ul style="list-style-type: none"> <li>• Observations</li> <li>• Interviews</li> <li>• Field notes</li> </ul>   |
| Cheng (2005)       | Hong Kong      | HKCEE  | Mixed Methods: <ul style="list-style-type: none"> <li>• Questionnaires</li> <li>• Observations</li> <li>• Follow-up interviews</li> </ul>  |
| Wall (2005)        | Sri Lanka      | O-level English Examination                    | Mixed Methods: <ul style="list-style-type: none"> <li>• Individual and group interviews</li> <li>• Questionnaires to teachers and teacher advisors</li> <li>• Material and test analysis</li> <li>• Observation</li> </ul> |
| Shih (2006)        | Taiwan         | General English Proficiency Test (GEPT)        | Qualitative: <ul style="list-style-type: none"> <li>• Interviews</li> <li>• Observation</li> </ul>   |
| Turner (2006)      | Canada         | Provisional Exams                              | Mixed Methods: <ul style="list-style-type: none"> <li>• Questionnaires</li> </ul>  |
| Scott (2007)       | United Kingdom | Primary School Statutory testing               | Qualitative: <ul style="list-style-type: none"> <li>• Interview</li> <li>• Observations</li> </ul>   |
| Fox & Cheng (2007) | Canada         | Ontario Secondary School Literacy Test (OSSLT) | Qualitative: <ul style="list-style-type: none"> <li>• Group Discussions</li> </ul>   |

|                |                |  |  |
|----------------|----------------|--|--|
| Tsagari (2007) | Hungary        | FCE examination<br>(Cambridge/ESL)                                 | Qualitative:<br><ul style="list-style-type: none"> <li>• Textbook Analysis</li> </ul>  |
| Green (2007)   | United Kingdom | IELTS writing component  | Mixed methods:<br><ul style="list-style-type: none"> <li>• Observations</li> <li>• Analysis of test components and test documents</li> <li>• Individual interviews</li> <li>• Focus group interviews</li> <li>• Questionnaires</li> </ul>                              |
| Mullen (2009)  | Canada         | Test of English for international communication                    | Mixed Methods:<br><ul style="list-style-type: none"> <li>• Questionnaires</li> <li>• Interviews</li> </ul>   |
| Tan (2009)     | Malaysia       | Bilingual high stakes secondary exit examination                   | Mixed Methods:<br><ul style="list-style-type: none"> <li>• Field notes,</li> <li>• Observations</li> <li>• Interviews</li> </ul>   |
| Turner (2009)  | Canada         | High stakes Quebec provincial secondary school exam                | Mixed Methods:<br><ul style="list-style-type: none"> <li>• MELS Exam Task Characteristic List (MTCs)</li> <li>• Semi-structured Interviews Classroom Observation</li> <li>• Guide (and field notes)</li> <li>• MTC Chart</li> <li>• Post-Observation Chats.</li> </ul> |
| Baker (2010)   | Canada         | English proficiency assessment for teacher certification in Quebec | Mixed methods:<br><ul style="list-style-type: none"> <li>• Interviews</li> <li>• Observations</li> <li>• Questionnaire</li> </ul>  |
| Wang (2010)    | China          | College English Test (CET)   | Mixed methods:<br>Questionnaires<br>In-depth case studies  |

|                             |           |  |   |
|-----------------------------|-----------|--|---|
| Cheng, Andrews & Yu (2011). | Hong Kong | School Based Assessment (SBA) in HKCEE | Quantitative: <ul style="list-style-type: none"><li>• Surveys</li></ul>                             |
| Erfani (2013)               | Iran      | IELTS & TOFEL iBt                      | Mixed methods <ul style="list-style-type: none"><li>• Questionnaires</li><li>• Interviews</li></ul> |

**Appendix D: Summary of Factors Identified by Empirical Studies as Affecting the Degree and Kinds of Washback (Source, Spratte, 2005, p. 29).**

| Teacher-related factors  | Resource, the school, the exam  |
|--|---|
| <p><i>Teacher beliefs about:</i></p> <ul style="list-style-type: none"> <li>● the reliability and fairness of the exam</li> <li>● what constitute effective teaching methods</li> <li>● how much the exam contravenes their current teaching practices</li> <li>● the stakes and usefulness of the exam</li> <li>● their teaching philosophy</li> <li>● about the relationship between the exam and the textbook</li> <li>● their students' beliefs</li> </ul> <p><i>Teachers' attitudes towards:</i></p> <ul style="list-style-type: none"> <li>● the exam</li> <li>● preparation of materials for exam classes</li> <li>● lesson preparation for exam classes</li> </ul> <p><i>Teachers' education and training:</i></p> <ul style="list-style-type: none"> <li>● Teachers' own education and educational experience</li> <li>● the amount of general methodological training they have received</li> <li>● training in teaching towards specific exams and in how to use exam-related textbooks</li> <li>● access to and familiarity with exam support materials such as exam specifications</li> <li>● understanding of the exam's rationale or philosophy.</li> </ul> <p><i>Other:</i></p> <ul style="list-style-type: none"> <li>● personality</li> <li>● willingness to innovate</li> </ul> | <p><i>Resources:</i></p> <ul style="list-style-type: none"> <li>● the availability of customised materials and exam support materials such as exam specifications</li> <li>● the types of textbooks available</li> </ul> <p><i>The school:</i></p> <ul style="list-style-type: none"> <li>● its atmosphere</li> <li>● how much the administrators put pressure on teachers to achieve results</li> <li>● the amount of time and number of students allocated to exam classes</li> <li>● cultural factors such as learning traditions</li> </ul> <p><i>The exam:</i></p> <ul style="list-style-type: none"> <li>● its proximity</li> <li>● its stakes</li> <li>● the status of the language it tests</li> <li>● its purpose</li> <li>● the formats it employs</li> <li>● the weighting of individual papers</li> <li>● when the exam was introduced</li> <li>● how familiar the exam is to teachers</li> </ul> |



8. If you were asked to identify what language teaching approach you are a proponent of? Would you say:

**Traditional grammar-based**  
**Post- communicative**  
**Other**

**Communicative**  
**Task-based teaching**  
**Please specify:**

**In the box below, please explain why you answered question 14 in the way you did?**

9. Is your classroom?

**Teacher -centered**

**Student-centered**

10. How do you find the implementation of new SECEE curriculum in your classroom?

**Easy**

**Moderately Difficult**

**Very Difficult**

**Impossible**

11. What obstacles have you encountered when implementing new SECEE curriculum?

(You can circle more than one if you wish).

**Culture**

**Students'  
Attitudes**

**Classroom  
Norms**

**Lack of  
Facilities**

**Institutional  
Regulations**

**Other**

If you would like to further clarify. Please feel free to do so in the box below:

12. Do you find there is a clash between teacher-student perceptions of learning within the classes you are currently delivering?

**Strongly agree      Agree      Strongly disagree      disagree**

13. Based on your experience are there any important aspects that seem to be missing in your schools that may help your Grade 12 teaching?

14. If you had a chance to improve your school or educational system, what aspect(s) would you focus on? Please be as explicit as possible?

**Thank you for your participation it is mostly appreciated**

شكرا

## Appendix F: Ethics Clearance Letter



Office of Research Ethics and Compliance  
 5110 Human Computer Interaction Bldg | 1125 Colonel By Drive  
 | Ottawa, Ontario K1S 5B6  
 613-520-2600 Ext: 2517  
[ethics@carleton.ca](mailto:ethics@carleton.ca)

### CERTIFICATION OF INSTITUTIONAL ETHICS CLEARANCE

The Carleton University Research Ethics Board-A (CUREB-A) at Carleton University has renewed ethics approval for the research project detailed below. CUREB-A is constituted and operates in compliance with the *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans* (TCPS2).

**Title:** The Washback of the New Secondary Education Certificate Examination of English (SECEE) in the Libyan EFL Classroom: A Multi-Methods Study [Nwara Abdulhamid]

**Protocol #:** 106098

**Project Team Members:** Nwara Abdulhamid (Primary Investigator)  
 Mr. Janna Fox (Research Supervisor)

**Department and Institution:** Faculty of Arts and Social Sciences\Linguistics and Language Studies (School of), Carleton University

**Funding Source (If applicable):**

Effective: January 15, 2018

Expires: January 31, 2019.

#### Restrictions:

This certification is subject to the following conditions:

1. Clearance is granted only for the research and purposes described in the application.
2. Any modification to the approved research must be submitted to CUREB-A. All changes must be approved prior to the continuance of the research.
3. An Annual Application for the renewal of ethics clearance must be submitted and cleared by the above date. Failure to submit the Annual Status Report will result in the closure of the file. If funding is associated, funds will be frozen.

4. A closure request must be sent to CUREB-A when the research is complete or terminated.
5. Should any participant suffer adversely from their participation in the project you are required to report the matter to CUREB-A.
6. It is the responsibility of the student to notify their supervisor of any adverse events, changes to their application, or requests to renew/close the protocol.
7. Failure to conduct the research in accordance with the principles of the *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans 2nd edition* and the *Carleton University Policies and Procedures for the Ethical Conduct of Research* may result in the suspension or termination of the research project.

Upon reasonable request, it is the policy of CUREB, for cleared protocols, to release the name of the PI, the title of the project, and the date of clearance and any renewal(s).

Please email the Research Compliance Coordinators at [ethics@carleton.ca](mailto:ethics@carleton.ca) if you have any questions or if you require a clearance certificate with a signature.

**CLEARED BY:**

**Date: January 15, 2018**



Andy Adler, PhD, Chair, CUREB-A



Bernadette Campbell, PhD, Vice-Chair, CUREB-A

### Appendix G: Student Questionnaire

Please circle the answer that is applicable to you

- |  |             |                  |           |
|--|-------------|------------------|-----------|
| 1. What is your gender?  | <b>Male</b> | <b>Female</b>    |           |
| 2. Have you taken the Secondary Education Certificate Examination of English (SECEE) before? | <b>Yes</b>  | <b>No</b>        |           |
| 3. Has somebody in your family taken the SECEE before?                                       | <b>Yes</b>  | <b>No</b>        |           |
| 4. Do you enjoy learning English?  | <b>Yes</b>  | <b>Somewhat</b>  | <b>No</b> |
| 5. Do you enjoy reading in English?  | <b>Yes</b>  | <b>Somewhat</b>  | <b>No</b> |
| 6. Do you enjoy speaking English?  | <b>Yes</b>  | <b>Somewhat</b>  | <b>No</b> |
| 7. Do you enjoy writing in English   | <b>Yes</b>  | <b>Somewhat</b>  | <b>No</b> |
| 8. Do you enjoy listening to English?  | <b>Yes</b>  | <b>Somewhat</b>  | <b>No</b> |
| 9. I am worried about my upcoming SECEE?   | <b>Yes</b>  | <b>Somewhat</b>  | <b>No</b> |
| 10. My parents are helping me prepare for the upcoming SECEE?                                | <b>Yes</b>  | <b>Somewhat</b>  | <b>No</b> |
| 11. My teacher is doing a lot of test preparation in the classroom?                          | <b>Yes</b>  | <b>Somewhat</b>  | <b>No</b> |
| 12. I expect my teacher to fully prepare me for the upcoming SECEE                           | <b>Yes</b>  | <b>Some what</b> | <b>No</b> |
| 13. I feel that I have control of the SECEE content  | <b>Yes</b>  | <b>Somewhat</b>  | <b>No</b> |

14. I am taking tutorial classes outside of school to help me prepare for the upcoming SECEE? **Yes** **Some what** **No**
15. I am doing a lot of test preparation outside of class time? **Yes** **Some what** **No**

**Thank you-** شكرا

## **Appendix H: Teacher Interview Questions**

Libyan Teacher Interview Questions (adapted from Abdulhamid, 2011, 2013)

### **Your current experiences in teaching Grade 12 SECEE Curriculum:**

#### **Think about classes you are currently teaching:**

1. How is it going? What is good and bad about how it is progressing? Are you finding any difficulties, challenges or “hotspots”?

#### **Think about a recent lesson/class:**

1. What were the objectives?
2. How did it go? What was good or bad about how it went? Were there any difficulties, challenges or hotspots?

### **Now we are going to the idea of testing and how the new SECEE and its relation to teachers and teaching, learners and learning.**

1. How does the test influence what you are teaching?
2. How does the test influence how you teach?
3. What in your view are the reasons behind the change in the current examination format and content?
4. What can you tell me about the test format?
5. Do you do any necessary extra work or find pressure in teaching towards the new SECEE?
6. Were there any changes in your teaching methodology after the introduction of the SECEE?
7. What about possible challenges whilst teaching the new SECEE curriculum?
8. What do you think about the influence of the SECEE on learners and learning?
9. What difficulties do you believe your students have with the new SECEE curriculum? What about the examination?
10. How do you think the new SECEE measures students’ language ability?
11. Do you consider the new SECEE examination a fair form of assessment?
12. Do you feel there is any teaching reform from which your school might profit?

## Appendix I: Grade 12 Content Objectives

|  |  |
|--|--|
| <p><b>Reading</b></p> <ol style="list-style-type: none"> <li>1) To develop skills in predicting the content of a text, including vocabulary.</li> <li>2) To read for detail.</li> <li>3) Taking notes</li> <li>4) To develop skills in assessing whether a writer is certain or uncertain</li> <li>5) To practise scanning a text.</li> <li>6) To practice Scanning for specific information.</li> <li>7) To work out the meaning of phrases from context.</li> <li>8) To develop pre-reading skills.</li> <li>9) To read and make notes</li> <li>10) To work out the meaning of phrases from context</li> <li>11) To develop the skill of explaining implicit meaning.</li> <li>12) To read a text and prepare to tell others about it.</li> <li>13) To retell a story, explaining, rephrasing and repeating if asked to do so.</li> <li>14) To express and justify opinions.</li> <li>15) Reading to retell information</li> <li>16) To practice reading and Identifying topic sentences</li> <li>17) To practice reading for global meaning</li> <li>18) Reading for specific information</li> <li>19) Understanding gist</li> <li>20) To raise awareness of different styles of writing.</li> <li>21) To practise reading for specific information.</li> <li>22) To practise reading fiction.</li> </ol> |  |
|--|--|

|   |  |
|---|--|
| <p>23) To develop the skill of understanding inferred meaning.</p> <p>24) To practice finding and interpreting topic sentences.</p> <p>25) To develop the skill of recognizing characterization and differences in style.</p> <p>26) To practise reading for detailed information.</p> <p>27) To practise scanning for vocabulary items.</p> <p>28) To deduce meaning from usage in a text.</p> <p>29) To introduce the meaning of verbs for reporting speech</p> <p>30) To raise awareness of collocations (prepositions that follow particular verbs).</p> <p>31) To raise awareness of reference words.</p> <p>32) To develop the skill of explaining meaning in different words.</p> <p>33) To raise awareness of the importance of English around the world.</p> |  |
| <p><b>1) Vocabulary</b></p> <p>1) To raise awareness of adverbs used to show degrees of certainty.</p> <p>2) To practice adjective + preposition combinations.</p> <p>3) To raise awareness of ways of referring to points of time in the future.</p> <p>4) To practise using <i>by</i>, <i>until</i> and other words and phrases used to refer to the future.</p> <p>5) To introduce vocabulary relating to success and failure.</p> <p>6) To introduce positive and negative connotations of vocabulary.</p>  |  |

|  |  |
|--|--|
| <p>7) To practice using new language in context.</p> <p>8) To practice forms of verbs make, do, and have.</p> <p>9) To discuss similarities and differences of meaning between vocabulary items.</p> <p>10) To introduce words and phrases for connecting sentences.</p> <p>11) To raise awareness of the gerund used as a noun.</p> <p>12) To practise using adjectives ending <i>-ing</i> and <i>-ed</i>.</p> <p>13) To discuss similarities and differences of meaning between vocabulary items.</p> <p>14) To introduce words and phrases for connecting sentences.</p> <p>15) To raise awareness of the formation and spelling of verbs and nouns.</p>  |  |
| <p><b>2) Grammar</b></p> <p>1) To review and practice subject and object questions</p> <p>2) Talking about the past with <i>must</i>, <i>may</i>, <i>might</i> and <i>can't</i>.</p> <p>3) To practice sentence patterns with <i>so</i>, <i>enough</i> and <i>too</i>.</p> <p>4) To practice the usual order of adjectives before a noun</p> <p>5) To practice interpreting writer's opinion.</p> <p>6) To practice using <i>must</i>, <i>may</i>, <i>might</i> and <i>can't</i> + have + pp to speculate about the past.</p> <p>7) To raise awareness of the future perfect and future continuous.</p> <p>8) To practise making predictions about specific times in the future.</p> <p>9) To raise awareness of formal English used by journalists.</p> |  |

|  |  |
|--|--|
| <p>10) To practise using <i>be</i> + infinitive to make statements about the future.</p> <p>11) To practise reading and writing newspaper headlines.</p> <p>12) To practice if condition type 3.</p> <p>13) To practise the future in the past: <i>was going to</i>.</p> <p>14) To raise awareness of the various uses of the infinitive.</p> <p>15) To introduce adjectives, verbs, nouns and question words followed by the infinitive.</p> <p>16) To review and extend awareness of verbs followed by <i>-ing</i> or the infinitive.</p> <p>17) To practise using <i>-ing</i> or the infinitive after specific verbs.</p> <p>18) To recognize written grammar mistakes.</p> <p>19) To analyze grammatical usage in context.</p> <p>20) To introduce patterns in reported speech.</p> <p>21) To use Verbs for reporting speech.</p> <p>22) To practice Time phrases and questions in reported speech.</p> <p>23) To practice using the passive in sentences in present and past sentences and with modal verbs.</p> <p>24) To recognise written grammar mistakes.</p> <p>25) To practise using the passive with continuous tenses.</p> <p>26) To raise awareness of <i>have</i> + object + past participle with a passive meaning.</p> |  |
| <p><b>3) Speaking</b></p> <p>1) Solving puzzles and responding to suggestions</p>  |  |

- 2) To practise speculating about puzzles using language from the unit.
- 3) To respond to suggestions politely.
- 4) To practise: *You could be right./That's one possibility, but.../That's a good idea.*
- 5) To talk about examples of extreme weather.
- 6) To prepare orally for the Writing lesson.
- 7) To practise spoken fluency in describing and narrating.
- 8) To practise sequencing expressions: *after that, in the end.*
- 9) To tell a story from pictures.
- 10) To prepare orally for the writing exercises.
- 11) To discuss the images and language used in persuasive posters and leaflets.
- 12) To practise giving advice.
- 13) To practise *You should (n't)/It's a good idea to/It's important (not) to/It's best (not) to.*
- 14) Telling a news story
- 15) Giving advice.
- 16) Telling a story from pictures.
- 17) Talking about books.
- 18) To exchange information in order to find inconsistencies.
- 19) To give and acknowledge compliments
- 20) To write a book review.
- 21) Giving instructions.
- 22) Giving opinions and comparing English with Arabic.

|  |  |
|--|--|
| <p>23) To discuss the images and language used in persuasive posters and leaflets.</p> <p>24) To practise describing procedures.</p> <p>25) To practise giving precise instructions.</p> <p>26) To practise <i>You should/Don't/Make sure you (don't)/Be careful (not) to/Always/Never.</i></p>  |  |
| <p><b>4) Writing</b></p> <ol style="list-style-type: none"> <li>1) To raise awareness of words and phrases used to introduce points of view.</li> <li>2) To write an article about a mysterious place or event using language from the unit.</li> <li>3) Identifying styles of writing.</li> <li>4) Presenting different points of view.</li> <li>5) To write newspaper report about an extreme weather event</li> <li>6) To practice paragraph structure.</li> <li>7) To practice topic sentences.</li> <li>8) To check written work for errors.</li> <li>9) To raise awareness of the persuasive language used in leaflets.</li> <li>10) To write a leaflet giving advice. To write a text based on discussion.</li> <li>11) To practise narrative cohesion.</li> <li>12) To practise recycled language from the Unit: <i>was going to; wish; conditionals.</i></li> <li>13) Leaflets giving advice.</li> <li>14) Writing a story.</li> <li>15) To raise awareness of the structure and language of reviews.</li> <li>16) To write a book review.</li> </ol> |  |

|  |  |
|--|--|
| <p>17) Longer sentences.</p> <p>18) To raise awareness of the language of instructions.</p> <p>19) To practise writing clear and precise instructions.</p> <p>20) Comparing and contrasting.</p> <p>21) To rewrite a story, making improvements in content and style.</p> <p>22) To produce a short piece of creative writing.</p> <p>23) To introduce paragraphing, reported speech, and words and phrases for connecting sentences.</p>  |  |
| <p><b>5) Listening</b></p> <p>1) Listening for key information.</p> <p>2) To predict the topic of a conversation.</p> <p>3) To practise listening for specific details.</p> <p>4) To raise awareness of rising and falling intonation to signal certainty and uncertainty in question tags.</p> <p>5) To listen to weather forecasts and understand important points.</p> <p>6) To pronounce and recognize key vocabulary.</p> <p>7) To practise listening for specific details.</p> <p>8) To listen to a presentation giving advice.</p> <p>9) To develop pre-listening skills.</p> <p>10) To practise listening for specific information.</p> <p>11) To listen to complete notes.</p> <p>12) To raise awareness of contrastive stress in sentences.</p> <p>13) To practise predicting the content of a conversation.</p> <p>14) To raise awareness of the pronunciation of consonant clusters.</p> <p>15) Listening to a weather forecast</p> <p>16) Listening for specific details and contrastive stress</p> |  |

|   |  |
|---|--|
| <p>17) Understanding information and instructions.</p> <p>18) Predicting content and listening for gist.</p> <p>19) To practise predicting the content of a conversation.</p> <p>20) To practise listening to assess the function of a conversation.</p> <p>21) To summarise the content of a conversation.</p> <p>22) To practice functional language.</p> <p>23) To practise listening to information and instructions.</p> <p>24) To practise distinguishing between vowel sounds.</p> |  |
|---|--|

### **Appendix J: Policy Maker Interview Questions**

1. What are the objectives of the new Secondary Education Certificate Examination of English curriculum?
2. What are the main reasons behind the current curricular reform policy?
3. What is the intended washback of the new SECEE on the Libyan EFL classroom?
4. What does the new SECEE intend to measure?

### **Appendix K: Test Developer Interview Questions**

1. Could you tell me about your experience and involvement with test development?
2. What are the objectives of the revised Secondary Education Certificate Examination of English curriculum?
3. What do you think are the main reasons behind the current curricular changes?
4. What does the revised SECEE intend to measure?
5. Could you explain to me how the test SECEE are developed?
6. How were the general test specifications formulated? and were the articulated test specifications built on the set Libyan EFL standards?
7. What was logic model that guided test design and item development?
8. How have the test items been validated?
9. Have you received any training in mental testing and statistical methods?
10. Do you think the examination tests the four language skills? If not, could you tell me what does it test and why?
11. In an ideal scenario, a test would measure students' different levels of thinking, for example, it will have a balanced combination of higher and lower order thinking skills. What different levels of thinking do you think the SECEE is measuring?
12. In what does the revised SECEE ensure that meaningful aspects of students thinking and learning are captured?
13. Would you say this is a language proficiency test, or an achievement test?
14. What do you think about the degree of alignment between the Libyan EFL standards and the examination? Could you explain to me how they align?
15. What would come to your mind? if I said that there is research evidence documenting that this SECEE is not aligned with the Libyan EFL standards.
16. Would you say that the revised SECEE is a fair form of assessment?

## Appendix L: World Language Cognitive Rigor Matrix (CRM, 2015) for the EFL descriptions of the depth of knowledge (DOK)

### World Language Cognitive Rigor Matrix

[Based on the Hess Cognitive Rigor Matrix (2005, 2009) & Hess' DOK supports for English Language Learners (2013)]

| World Language Practices & Modes of Communication                                    | Webb's DOK Level 1<br>Recall & Reproduction<br><i>(Having the knowledge/language required; don't need to "figure it out")</i>  | Webb's DOK Level 2<br>Skills & Concepts<br><i>(Making connections among skills/concepts or decisions - e.g., about approach, general message/concepts)</i>  | Webb's DOK Level 3<br>Strategic Thinking/ Reasoning<br><i>(Complex &amp; Abstract; Exploring multiple solution paths; Justifying with evidence)</i>   | Webb's DOK Level 4<br>Extended Thinking<br><i>(Relating /developing complex ideas using multi-sources and evidence)</i>  |
|--|--|---|---|--|
| <b>Memorize &amp; Recall</b>   | <ul style="list-style-type: none"> <li>• Reproduce/recall/repeat vocabulary, grammar rules, facts, definitions, dictated statements, etc.</li> <li>• Describe cultural conventions</li> <li>• Recite in sequence (e.g., alphabet, counting, songs, rhymes)</li> </ul>  | <p>Use these World Language CRM curricular examples for designing most language and communication assignments or assessment tasks.</p>  |   |  |
| <b>Interpersonal Communication</b><br><br><b>Understand, Perceive, &amp; Respond</b> | <ul style="list-style-type: none"> <li>• Understand simple, familiar messages in social settings</li> <li>• Identify everyday objects</li> <li>• Follow simple oral directions or written procedures (recipe, etc.)</li> <li>• Convey simple messages, express feelings (e.g., I'm sad because...)</li> <li>• Ask/answer literal questions after reading, listening, or viewing</li> </ul> | <ul style="list-style-type: none"> <li>• Explain how or why alternative responses may be correct (where do you live?) for different situations</li> <li>• Carry on a short conversation using familiar vocabulary and grammar</li> <li>• Paraphrase/summarize/retell what was said, read, viewed (with cues)</li> <li>• Make logical predictions (e.g., what might happen next...); describe event</li> </ul>         | <ul style="list-style-type: none"> <li>• Prepare for an interview or develop survey on topic of interest anticipating audience questions/ possible responses</li> <li>• Initiate &amp; extend a conversation about an unfamiliar topic, appropriately using language mechanics/tense throughout</li> <li>• Create a theme-based photo essay</li> <li>• Justify interpretation of purpose or tone (in media message, photo essay, etc.)</li> </ul> | <ul style="list-style-type: none"> <li>• Carry on an extended conversation responding appropriately to multiple speakers (e.g., using multiple tenses, asking and answering, elaborating on ideas, raising questions)</li> <li>• Deepen knowledge of a topic using multiple (oral, visual, textual) sources for an informational communication (e.g., "by the numbers" infographic)</li> </ul>   |
| <b>Interpret &amp; Apply</b>   | <ul style="list-style-type: none"> <li>• Match vocabulary (e.g., picture-word; synonyms); locate details</li> <li>• Apply a spelling or grammar rule (e.g., conjugate a verb, make plural)</li> <li>• Use resources to translate literally</li> <li>• Use nouns/verbs in familiar contexts</li> </ul>  | <ul style="list-style-type: none"> <li>• Infer and explain meaning using context, cognates, or structure in a familiar situation</li> <li>• Translate to identify use of non-literal/figurative/idiomatic language</li> <li>• Sequence events for given text/visual</li> </ul>  | <ul style="list-style-type: none"> <li>• Explain inferences or colloquial expressions using supporting evidence</li> <li>• Interpret symbolic/abstract meaning (from music, video, reading, art, etc.)</li> <li>• Interpret idiomatic/ figurative language in context (poem, song lyric, media, etc.)</li> </ul>  | <ul style="list-style-type: none"> <li>• Make and justify conclusions based on 2+ ads for the same product or two political cartoons about the same event or person</li> <li>• Write/draw/perform in the style of a known author/artist/cartoonist</li> </ul>  |
| <b>Compare, Analyze, Critique/Evaluate, &amp; Reflect</b>                            | <ul style="list-style-type: none"> <li>• Edit a sentence/phrase</li> <li>• Select appropriate word/phrase for intended meaning</li> <li>• Answer what/when/where questions using a source (map, calendar, schedule, visual, photo)</li> <li>• Connect words/phrases between languages (origins, meanings, etc.)</li> </ul>   | <ul style="list-style-type: none"> <li>• Categorize/ compare (objects, foods, tools, people, etc.) using oral/physical/textual stimuli</li> <li>• Self-correct when speaking or reading</li> <li>• Evaluate message or cultural nuances (e.g., gestures, language) using listening and observational skills</li> </ul>  | <ul style="list-style-type: none"> <li>• Evaluate &amp; correct inaccuracy of a message - print or non-print text (e.g., facts, sequence, cultural nuances)</li> <li>• Support an opinion/argument/ disagreement with evidence, reasoning</li> <li>• Determine if source can/cannot answer specific questions &amp; why (e.g., websites)</li> </ul>   | <ul style="list-style-type: none"> <li>• Critique authentic literature/arts/ historical events from multiple sources: authors/perspectives/time periods</li> <li>• Evaluate relevancy, accuracy, &amp; completeness of information</li> <li>• Keep a journal and use it to reflect on/evaluate personal progress</li> </ul>  |
| <b>Presentational Communication</b><br><br><b>Produce or Create</b>                  | <ul style="list-style-type: none"> <li>• Represent vocabulary/common phrases in pictures, symbols, visuals, gestures, pantomime</li> <li>• Brainstorm related words, ideas, images, possible responses</li> <li>• Label information on a diagram, map, visual</li> <li>• Tell/select phrases as thumbnail sketch for a narrative text/ story line</li> </ul>                               | <ul style="list-style-type: none"> <li>• Perform a memorized dialog</li> <li>• Choose which tense to use in a less familiar context</li> <li>• Create an ABC book connecting entries by central /organizing topic (e.g., animals, foods)</li> <li>• Create text messages or description (narration/voice over) for a visual stimuli or "muted" video scene</li> <li>• Make/label a time line of key events</li> </ul> | <ul style="list-style-type: none"> <li>• Develop a vocabulary-based game to teach about geography, culture, etc.</li> <li>• Develop a new scene/ending, consistent with the original text</li> <li>• Create or perform a dialog based on visual stimuli or a current or cultural event (integrating academic vocabulary)</li> <li>• Co-plan website/event highlighting target culture (foods, traditions, places to visit)</li> </ul>             | <ul style="list-style-type: none"> <li>• Produce an 'old' idea in a new way (e.g., multi-media, podcast)</li> <li>• Integrate ideas from several sources</li> <li>• Research a topic with evidence pro-con for debate/ essay/cartoon</li> <li>• Research and present performance/ presentation using multiple sources</li> <li>• Design a theme-based café, including the menu, location/décor and develop an ad for targeted clientele</li> </ul> |

**Appendix M: Coding Matrix I**

## Coding Matrix II

**Please assign each listed content objective with the best corresponding depth-of-knowledge level. If you have difficulty deciding between two levels for a content objective (e.g., between a rating of level 1 or 2), you are advised to choose the higher of the two levels.**

| Content objective   | DOK<br>Level |
|---|--------------|
| <b>6) Reading</b>   |              |
| 1. To develop skills in predicting the content of a text, including vocabulary. |              |
| 2. To read for detail.  |              |
| 3. Taking notes   |              |
| 4. To develop skills in assessing whether a writer is certain or uncertain      |              |
| 5. To practise scanning a text.   |              |
| 6. To practice Scanning for specific information.                               |              |
| 7. To work out the meaning of phrases from context.                             |              |
| 8. To develop pre-reading skills.   |              |
| 9. To read and make notes   |              |
| 10. To work out the meaning of phrases from context                             |              |

|  |  |
|--|--|
| 11. To develop the skill of explaining implicit meaning.                           |  |
| 12. To read a text and prepare to tell others about it.                            |  |
| 13. To retell a story, explaining, rephrasing and repeating if asked to do so.     |  |
| 14. To express and justify opinions.   |  |
| 15. Reading to retell information  |  |
| 16. To practice reading and Identifying topic sentences                            |  |
| 17. To practice reading for global meaning   |  |
| 18. Reading for specific information   |  |
| 19. Understanding gist   |  |
| 20. To raise awareness of different styles of writing.                             |  |
| 21. To practise reading for specific information.                                  |  |
| 22. To practise reading fiction.   |  |
| 23. To develop the skill of understanding inferred meaning.                        |  |
| 24. To practice finding and interpreting topic sentences.                          |  |
| 25. To develop the skill of recognizing characterization and differences in style. |  |
| 26. To practise reading for detailed information.                                  |  |
| 27. To practise scanning for vocabulary items.                                     |  |
| 28. To deduce meaning from usage in a text.  |  |
| 29. To introduce the meaning of verbs for reporting speech                         |  |

|  |  |
|--|--|
| 30. To raise awareness of collocations (prepositions that follow particular verbs).                    |  |
| 31. To raise awareness of reference words.   |  |
| 32. To develop the skill of explaining meaning in different words.                                     |  |
| 33. To raise awareness of the importance of English around the world.                                  |  |
| <b>7) Vocabulary</b>   |  |
| 1. To raise awareness of adverbs used to show degrees of certainty.                                    |  |
| 2. To practice adjective + preposition combinations.   |  |
| 3. To raise awareness of ways of referring to points of time in the future.                            |  |
| 4. To practise using <i>by</i> , <i>until</i> and other words and phrases used to refer to the future. |  |
| 5. To introduce vocabulary relating to success and failure.  |  |
| 6. To introduce positive and negative connotations of vocabulary.                                      |  |
| 7. To practice using new language in context.  |  |
| 8. To practice forms of verbs make, do, and have.  |  |
| 9. To discuss similarities and differences of meaning between vocabulary items.                        |  |
| 10. To introduce words and phrases for connecting sentences.   |  |
| 11. To raise awareness of the gerund used as a noun.   |  |
| 12. To practise using adjectives ending <i>-ing</i> and <i>-ed</i> .                                   |  |
| 13. To discuss similarities and differences of meaning between vocabulary items.                       |  |
| 14. To introduce words and phrases for connecting sentences.   |  |

|  |  |
|--|--|
| 15. To raise awareness of the formation and spelling of verbs and nouns.                               |  |
| <b>8) Grammar</b>  |  |
| 1. To review and practice subject and object questions   |  |
| 2. Talking about the past with <i>must, may, might</i> and <i>can't</i> .                              |  |
| 3. To practice sentence patterns with <i>so, enough</i> and <i>too</i> .                               |  |
| 4. To practice the usual order of adjectives before a noun   |  |
| 5. To practice interpreting writer's opinion.  |  |
| 6. To practice using <i>most, may, might</i> and <i>can't + have + pp</i> to speculate about the past. |  |
| 7. To raise awareness of the future perfect and future continuous.                                     |  |
| 8. To practise making predictions about specific times in the future.                                  |  |
| 9. To raise awareness of formal English used by journalists.   |  |
| 10. To practise using <i>be + infinitive</i> to make statements about the future.                      |  |
| 11. To practise reading and writing newspaper headlines.   |  |
| 12. To practice if condition type 3.   |  |
| 13. To practise the future in the past: <i>was going to</i> .  |  |
| 14. To raise awareness of the various uses of the infinitive.  |  |
| 15. To introduce adjectives, verbs, nouns and question words followed by the infinitive.               |  |
| 16. To review and extend awareness of verbs followed by <i>-ing</i> or the infinitive.                 |  |
| 17. To practise using <i>-ing</i> or the infinitive after specific verbs.                              |  |

|  |  |
|--|--|
| 18. To recognize written grammar mistakes.   |  |
| 19. To analyze grammatical usage in context.   |  |
| 20. To introduce patterns in reported speech.  |  |
| 21. To use Verbs for reporting speech.   |  |
| 22. To practice Time phrases and questions in reported speech.                                     |  |
| 23. To practice using the passive in sentences in present and past sentences and with modal verbs. |  |
| 24. To recognise written grammar mistakes.   |  |
| 25. To practise using the passive with continuous tenses.  |  |
| 26. To raise awareness of <i>have</i> + object + past participle with a passive meaning.           |  |
| <b>9) Speaking</b>   |  |
| 1. Solving puzzles and responding to suggestions   |  |
| 2. To practise speculating about puzzles using language from the unit.                             |  |
| 3. To respond to suggestions politely.   |  |
| 4. To practise: <i>You could be right./That's one possibility, but.../That's a good idea.</i>      |  |
| 5. To talk about examples of extreme weather.  |  |
| 6. To prepare orally for the Writing lesson.   |  |
| 7. To practise spoken fluency in describing and narrating.   |  |
| 8. To practise sequencing expressions: <i>after that, in the end.</i>                              |  |
| 9. To tell a story from pictures.  |  |

|  |  |
|--|--|
| 10. To prepare orally for the writing exercises.   |  |
| 11. To discuss the images and language used in persuasive posters and leaflets.                        |  |
| 12. To practise giving advice.   |  |
| 13. To practise <i>You should (n't)/It's a good idea to/It's important (not) to/It's best (not)to.</i> |  |
| 14. Telling a news story   |  |
| 15. Giving advice.   |  |
| 16. Telling a story from pictures.   |  |
| 17. Talking about books.   |  |
| 18. To exchange information in order to find inconsistencies.  |  |
| 19. To give and acknowledge compliments  |  |
| 20. To write a book review.  |  |
| 21. Giving instructions.   |  |
| 22. Giving opinions and comparing English with Arabic.   |  |
| 23. To discuss the images and language used in persuasive posters and leaflets.                        |  |
| 24. To practise describing procedures.   |  |
| 25. To practise giving precise instructions.   |  |
| 26. To practise <i>You should/Don't/Make sure you (don't)/Be careful (not) to/Always/Never.</i>        |  |
| <b>10) Writing</b>   |  |
| 1. To raise awareness of words and phrases used to introduce points of view.                           |  |

|  |  |
|--|--|
| 2. To write an article about a mysterious place or event using language from the unit.     |  |
| 3. Identifying styles of writing.  |  |
| 4. Presenting different points of view.  |  |
| 5. To write newspaper report about an extreme weather event                                |  |
| 6. To practice paragraph structure.  |  |
| 7. To practice topic sentences.  |  |
| 8. To check written work for errors.   |  |
| 9. To raise awareness of the persuasive language used in leaflets.                         |  |
| 10. To write a leaflet giving advice. To write a text based on discussion.                 |  |
| 11. To practise narrative cohesion.  |  |
| 12. To practise recycled language from the Unit: <i>was going to; wish; conditionals</i> . |  |
| 13. Leaflets giving advice.  |  |
| 14. Writing a story.   |  |
| 15. To raise awareness of the structure and language of reviews.                           |  |
| 16. To write a book review.  |  |
| 17. Longer sentences.  |  |
| 18. To raise awareness of the language of instructions.                                    |  |
| 19. To practise writing clear and precise instructions.                                    |  |
| 20. Comparing and contrasting.   |  |

|  |  |
|--|--|
| 21. To rewrite a story, making improvements in content and style.  |  |
| 22. To produce a short piece of creative writing.  |  |
| 23. To introduce paragraphing, reported speech, and words and phrases for connecting sentences.              |  |
| <b>11) Listening</b>   |  |
| 1. Listening for key information.  |  |
| 2. To predict the topic of a conversation.   |  |
| 3. To practise listening for specific details.   |  |
| 4. To raise awareness of rising and falling intonation to signal certainty and uncertainty in question tags. |  |
| 5. To listen to weather forecasts and understand important points.   |  |
| 6. To pronounce and recognize key vocabulary.  |  |
| 7. To practise listening for specific details.   |  |
| 8. To listen to a presentation giving advice.  |  |
| 9. To develop pre-listening skills.  |  |
| 10. To practise listening for specific information.  |  |
| 11. To listen to complete notes.   |  |
| 12. To raise awareness of contrastive stress in sentences.   |  |
| 13. To practise predicting the content of a conversation.  |  |
| 14. To raise awareness of the pronunciation of consonant clusters.   |  |

|   |  |
|---|--|
| 15. Listening to a weather forecast                                 |  |
| 16. Listening for specific details and contrastive stress           |  |
| 17. Understanding information and instructions.                     |  |
| 18. Predicting content and listening for gist.                      |  |
| 19. To practise predicting the content of a conversation.           |  |
| 20. To practise listening to assess the function of a conversation. |  |
| 21. To summarise the content of a conversation.                     |  |
| 22. To practice functional language.                                |  |
| 23. To practise listening to information and instructions.          |  |
| 24. To practise distinguishing between vowel sounds.                |  |
|   |  |

## Appendix N: Coding Matrix II

### Coding Matrix II

Please assign each listed test item with the best corresponding DOK level and objective(s) for the sixty test items. If you have difficulty deciding between two objectives for an item, you are advised to choose the two most suitable objectives (please refer to the objective listing handout). Could you also kindly assign a possible learning outcome(s) to each test item.

| Test item # | DOK level | Objective(s) |
|-------------|-----------|--------------|
| 1.          |           |              |
| 2.          |           |              |
| 3.          |           |              |
| 4.          |           |              |
| 5.          |           |              |
| 6.          |           |              |
| 7.          |           |              |
| 8.          |           |              |
| 9.          |           |              |
| 10.         |           |              |
| 11.         |           |              |
| 12.         |           |              |
| 13.         |           |              |
| 14.         |           |              |
| 15.         |           |              |
| 16.         |           |              |
| 17.         |           |              |
| 18.         |           |              |
| 19.         |           |              |
| 20.         |           |              |
| 21.         |           |              |
| 22.         |           |              |
| 23.         |           |              |
| 24.         |           |              |
| 25.         |           |              |
| 26.         |           |              |
| 27.         |           |              |
| 28.         |           |              |
| 29.         |           |              |

|     |  |  |
|-----|--|--|
| 30. |  |  |
| 31. |  |  |
| 32. |  |  |
| 33. |  |  |
| 34. |  |  |
| 35. |  |  |
| 36. |  |  |
| 37. |  |  |
| 38. |  |  |
| 39. |  |  |
| 40. |  |  |
| 41. |  |  |
| 42. |  |  |
| 43. |  |  |
| 44. |  |  |
| 45. |  |  |
| 46. |  |  |
| 47. |  |  |
| 48. |  |  |
| 49. |  |  |
| 50. |  |  |
| 51. |  |  |
| 52. |  |  |
| 53. |  |  |
| 54. |  |  |
| 55. |  |  |
| 56. |  |  |
| 57. |  |  |
| 58. |  |  |
| 59. |  |  |
| 60. |  |  |

If you have any comments

.....

.....

.....

.....

.....

.....

# Unit 1

## Puzzles and mysteries

### Lessons 1 & 2: Reading: Predicting content

#### 1. Before you read [Lesson 1]

**A** Look at the photos on page 7. Then discuss these questions in pairs.

1. Which of the photos was taken from a plane?
2. What can you see in each photo?
3. How old do you think the lines in each photo are?
4. Who or what do you think made the lines?

**B** Circle the word in each pair which you think you will find in the text. Discuss your reasons with a partner.

1. desert / sea
2. trees / ground
3. straight / short
4. colour / shape
5. flat / mountainous
6. paths / road
7. sandy / stony
8. draw / write
9. aliens / human beings
10. uncertain / unlikely

#### 2. While you read

**A** Read the first two paragraphs of the text. Find the answers to these questions.

1. Where are the Nazca lines?
2. What size is the area covered by the pictures?
3. Why didn't people discover the pictures until the 1930s?
4. How old do scientists think the pictures are?

**B** Match each diagram to a sentence.

1. The lines form a shape.
2. The lines are randomly placed.
3. The lines are parallel.
4. The lines intersect.



## Reading

**C** Read the last two paragraphs. Write notes about each of the following in your notebook.

1. what the two mysteries of the Nazca lines are
2. theories about the mysteries
3. the writer's opinion about the theories

**D** Compare your answers with a partner and discuss these questions.

1. Which of the theories about the Nazca lines do you think is the most believable?
2. Which of the theories do you think is the least believable?
3. Have you got any more ideas about how the lines were made?

www.puzzlesandmysteries.com

# The mystery of the Nazca lines

When planes fly over the Peruvian desert about 400 kilometres south of Lima, the passengers look down and see large pictures on the ground far below. As well as pictures of birds and animals, they can see hundreds of perfectly straight lines many kilometres long. Some of the lines are parallel, some of them intersect, some combine to form a shape and some seem to be randomly placed. These lines and pictures cover a flat area 60 kilometres long and two kilometres wide.

They can be seen only from a plane high above the ground. In fact, they were not discovered until planes began to fly over the area in the 1930s. If you are on the ground, you see only narrow paths through the stony desert. Apparently, the people who made the lines were able to look at the ground from the air in some way. But they can't have had planes or helicopters. According to most scientists who have studied the pictures, they are nearly 2,000 years old.

Clearly, there are two big mysteries about the Nazca lines. The first is this: how were the lines and pictures made 2,000 years ago? As far as we know, nobody could see the pictures, so it must have been difficult to draw them. The second and greater mystery is: why did they do it? If nobody on Earth could see the results, why did they bother? The lines and pictures must have had an important purpose. What was that purpose?

Many people have tried to answer these questions. Some people say the markings can't have been made by ancient people. They say they might have been made by aliens who could see the pictures from their spaceships. Other people have suggested that the ancient people might have made hot air balloons from animal skins and that a master artist might have directed teams of workers from his balloon. But it is extremely unlikely that these explanations are true. It is more likely that the ancient people found a simpler way to make the markings without any need for spaceships, planes or balloons. The purpose of the markings was most likely religious. The people may have thought their gods would see the pictures from the sky. Perhaps they also used the lines as paths in religious ceremonies. However, nobody knows for sure. The 'How?' and 'Why?' of the Nazca lines will always be a mystery.




### 3. After you read [Lesson 2]

**A** Now do Exercises A to E on Workbook page 4.

### Lesson 3: Vocabulary: Certainty and uncertainty

- A** Look at the picture and read what the vet says. What do you think it is? Do you think the vet is sure what happens?



- B** Look at the two sentences below. What is Tarek more likely to do, succeed in his exams or go to university? Read the information in the language box to help you.

1. Tarek is clearly going to do well in his exams.
2. He is probably going to university next year.

apparently clearly likely unlikely definitely actually probably

#### Degrees of certainty – adverbs

We can show how certain or uncertain we are by using adverbs. If we are sure about something, we use *actually*, *clearly* or *definitely*. If we are less sure about something, we use *possibly*, *probably* or *apparently*.

We put these adverbs:

- after the verb to be

**Example:** He is clearly an intelligent boy.

- before other verbs

**Example:** He clearly works hard.

- between an auxiliary verb (*be*, *have*, *will*, *can*, *do*, etc.) and a main verb in positive sentences

**Example:** He has clearly studied hard all year.

- before an auxiliary verb in negative sentences

**Example:** He clearly didn't want to fail the exams.

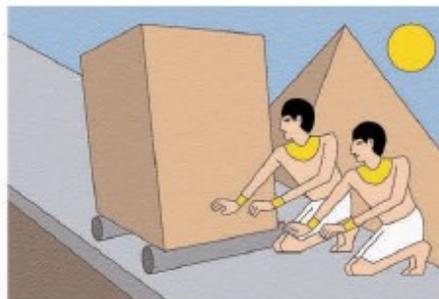
**Note:** *Actually*, *clearly*, *probably*, *possibly* and *apparently* can also be used at the beginning or end of a sentence.

**Example:** Clearly, he is the best student in the class.

- C** Now do Exercises A and B on Workbook page 5.

## Lesson 4: Grammar 1: Subject and object questions

**A** In pairs describe the pictures.



**B** Match the questions 1–8 to the answers a–h.

- |  |                          |   |
|--|--------------------------|---|
| 1. Who built the Great Pyramid?            | <input type="checkbox"/> | a) They rolled them on long pieces of wood. |
| 2. When did they build it?                 | <input type="checkbox"/> | b) About 30 years.                          |
| 3. How long did it take?                   | <input type="checkbox"/> | c) To make a tomb for the pharaoh.          |
| 4. How many people helped to build it?     | <input type="checkbox"/> | d) Probably from Aswan.                     |
| 5. What did they use to build the pyramid? | <input type="checkbox"/> | e) About 20,000.                            |
| 6. Where did the materials come from?      | <input type="checkbox"/> | f) Stone.                                   |
| 7. How did they transport them?            | <input type="checkbox"/> | g) The ancient Egyptians.                   |
| 8. Why did they build it?                  | <input type="checkbox"/> | h) 4,600 years ago.                         |

**C** Look carefully at the questions in Exercise B. How are questions 1 and 4 different from the other six questions? Check by reading the information below.

### Subject and object questions

In the question *Who built the Great Pyramid?* we want to find out information about the subject of the verb (*the Ancient Egyptians*). This type of question is sometimes called a subject question.

In the question *Why did they build it?* we already know the subject (*they*), and so we are asking about something else (*the reason why*). This type of question is sometimes called an object question.

We make subject questions without *do* or *did*. They usually begin with *who* or *what*.

Examples:

**Who** gave you my e-mail address? [Answer: Katie gave me your e-mail address.]

**What** makes him run? [Answer: Ambition makes him run.]

We use *do(n't)* or *did(n't)* in object questions in the present or simple past.

Examples:

**Where** did you get that scarf? [Answer: In Tripoli.]

**Why** does wood float? [Answer: Because it is less dense than water.]

**Why** didn't you open the door? [Answer: Because I couldn't find the key.]

**D** Now do Exercises A to D on Workbook pages 6–7

## Lesson 6: Speaking: Solving puzzles and responding to suggestions

### A Look at the pictures. What is the problem?



### B Read the conversation below. How do you think they solved the problem?

A: Well, the problem is that the lorry is stuck.

B: Yes, the bridge is too low and the driver can't get through.

A: In this picture, the boy has found a solution. What do you think it is?

B: The people might have pushed the lorry from the back.

A: That's one possibility, but I'm not sure it would have worked.

B: They might have cut off part of the bridge to make more room.

A: Hmm. I don't think that's very likely. It would have damaged the bridge.

B: Do you think the driver might have driven very fast towards the bridge to get through?

A: You could be right. But it might have been too dangerous.

B: Aha! I know how they must have solved the problem!



### C Read the language box about responding to suggestions.

#### Responding to suggestions

When someone makes a suggestion you don't completely agree with, there are some expressions you can use to be polite.

#### Examples:

*That's one possibility, but ...*

*I don't think that's very likely .../I think that's unlikely ...*

*You could be right.*

*That's a good idea, but ...*

### D Work in pairs. Read the puzzle below and talk about possible solutions. Use *must/might/can't + have + past participle* and adverbs from Lesson 3. Respond to your partner's suggestions.

A police chief was interviewing three candidates for a job in his department. To test their logic, he took a red marker pen and a black marker pen and told them, 'I am going to make either a red or black mark on each of your foreheads. At least one mark will be black. Using only your own logic, I want you to find out the colour of the mark on your own forehead. The first man to do this and give me an explanation of how he made his decision will get the job.'

He then blindfolded the candidates and put a black mark on each of their foreheads. After he removed the blindfolds, the three stared at each other for a few seconds, each seeing that the other two marks were black. Then one of the candidates said, 'I have a black mark'.

How did the candidate explain his decision?

### E Now do Exercise A on Workbook page 8.

## Lesson 7: Writing: Presenting different points of view

### 1. Preparation for writing

- A** Read the article and find four theories about the disappearance of the dinosaurs. Write brief notes about each one in your notebook.

## Was T-Rex killed by a tiny insect?

**A new theory has renewed scientific debate about exactly why dinosaurs disappeared from the face of the Earth 65 million years ago.**

According to George and Roberta Poinar from Oregon State University, tiny insects might have played an important role in wiping out the giant beasts. The husband-and-wife team have spent their lives studying the insect and plant life which is preserved in fossilized amber. They believe that a variety of insects may have spread infectious diseases or caused stomach problems which gradually made the dinosaurs die out.

In the 1980s, most people believed the theory of Professors Luis and Walter Alvarez as the most likely reason for the dinosaurs' extinction. In their view, the dinosaurs must have been killed by a giant asteroid hitting the Earth. More recently, a team of German scientists led by Peter

Schulte claimed that a series of volcanic eruptions were the cause of the dinosaurs' disappearance. They believed that these eruptions released toxic elements like cadmium and nickel into the atmosphere. Others have said that the planet may have been hit by a massive storm which killed off the dinosaurs.

In fact, none of these *sudden death* theories is convincing. The dinosaurs can't have disappeared so suddenly. Apparently, fossil evidence shows that extinction was a gradual process, which happened over millions of years.

The Poinars' theory is much more likely to be true. Actually, George and Roberta agree that



insects were probably just one factor in the disappearance of the dinosaurs. Climate change could also have contributed to this gradual process. Because dinosaurs were cold-blooded, they might not have survived increasingly cold temperatures. Perhaps it was more and more difficult for them to find food in the colder climate, experts argue.

We may never know exactly who or what killed the dinosaurs. But it seems that new ideas about this mystery will continue to fascinate future generations.

Adapted from: Science Daily, Jan 4, 2008, <http://www.sciencedaily.com/releases/2008/01/080103090702.htm>

- B** Look at the ways of introducing points of view in the Language box. Find and underline examples in the text.

#### Introducing points of view

- In fact, ...
- (Scientists) claim that ...
- According to ...
- In their view, ...
- (Experts) have put forward the idea that ...
- They believe that ...

- C** Choose a mysterious place or event from this unit or a mystery of your own. Do some research in a library or on the Internet. Make notes about different theories people have used to explain the mystery.

### 2. Writing

- A** Now do Exercise A on Workbook page 9.

## **Appendix P: Student Focus Group Questions**

### **Questions prior the SECEE**

1. Could you tell me about your experience learning English?
2. How comfortable are you with the new SECEE?
3. Have you had any challenges whilst learning and preparing for the test?
4. What kind and out-of-class English learning practices have you been doing?
5. Do you consider the new SECEE as a fair assessment?

### **Questions after the sitting the SECEE**

6. How did you feel during the examination?
7. How well did you deliver the test?
8. Were the test instructions clear?
9. Was the test format clear to you?
10. Were you aware of the weighting of test items and sections?
11. Were the test items clear?
12. Did you find any of the test items misleading?
13. Did you get a feeling that you did not now what exactly the items were testing?

## **Appendix Q: Extra Detail on Data Collection Procedures**

### **1. Teacher Questionnaire**

In particular, the teacher questionnaire comprised of two sections. The first part elicited demographic information and consisted of six items (four multiple choice items and two constructed response items) that covered gender, age range, educational background, years of EFL teaching, and years of Grade 12 EFL teaching. The information gave me a general picture about the teachers' background. The second section explored the teachers' beliefs about EFL teaching and learning and consisted of eight items; four multiple choice items, three constructed response items, and one 4-point Likert scale ranging from "strongly agree" to "strongly disagree". More specifically, it assesses teachers accounts of the revised secondary level curriculum and the rSECEE and their attitudes towards the revised testing system and its implementation into the Libyan education system. Space was allotted after each question to give the participants an idea of how much they were expected to write. Overall, the data obtained were in the form of items and extended written responses.

### **2. Observation**

My experience as a high school teacher and university instructor as well as the research literature about secondary level education and English language teaching in Libya have helped me build me an initial picture of what the Grade 12 EFL classroom dynamics would possibly look like. Thereafter, the use of observations helped to further my understanding of how Libyan EFL students are taught English within a classroom environment. More importantly, the observations offered me the chance to assess how teachers and their students interact within the current testing system. It also allowed me to witness how the system was affecting teachers' classroom behaviours, and therefore the washback direction and its intensity.

Before the classroom observations, I contacted the three teacher participants to advise that HM would only sit at the back of class to observe and take notes. It was also pointed out that HM was not evaluating their teaching. Furthermore, it was stressed that their participation was of great value to the Libyan education system as it provided policy makers with feedback from an important stakeholder group about the success of the new curricular implementation. It was also emphasized that the participants would be given pseudonyms, and the transcripts labelled with these pseudonyms, while interview transcripts were to be altered to remove any information that might lead to their identification.

Video recording of the 15 classes might have been useful but was not deemed necessary because along with the audio recording detailed notes were taken of the classes.

After each observation, the observational tools and the note taking observational matrix were completed by listening to the audio recording. Similar to Watanabe (2004), the observer's impressions and accounts of the observed lesson and field notes were recorded as narrative monographs in order to clearly recall the classroom situation and events of each class.

### **3. Interviews**

The interviews, which were conducted with the three Libyan teachers who had previously been observed, were employed because as noted by Denscombe (2010) they are a good method for eliciting data or participants' accounts, which are based on their priorities, opinions and ideas. The purpose of conducting the interviews was two-fold: 1) to elicit further information on the questionnaire responses and teachers' perceptions and themes by enabling the participants to expand on their ideas, explain their opinions, and identify what they regard as important; and 2) to gain a better understanding about the observed teachers' behaviours and classroom practices. The interviews were conducted after the results from the teacher questionnaire and observations became clear.

During the interviews, the teachers were asked questions that the researcher had prepared in advance. However, the researcher also asked other questions to probe points that came up during the interview. The interview questions (see Appendix H) were designed with the intention of exploring teachers' accounts and beliefs about the influence of the rSECEE on their teaching, and if they had made any changes in their approach to teaching as a result of the examination.

### **4. Focus Groups**

This study used the focus group interviews to elicit feedback from test-takers about their: background; perceptions of the rSECEE; experience of learning English; perceived impact of the rSECEE on their learning; challenges whilst learning and preparing for the test; and out-of-class English learning practices. In addition, the focus group elicited feedback regarding test transparency and issues related to test administration (for example, whether students felt exhausted during the test, or if they were able to deliver their best performance) because test-takers ought to be familiar with test tasks before the actual test is taken. In addition, Weir (2005) notes that the

level of test-takers' familiarity with test demands may affect the manner in which the test task is approached, and, consequently, affect test performance.

Focus group interviews with students were used as a triangulation tool for Phase II. I investigated the students' views of the rSECEE because I wanted to compare their accounts with those of their teachers, to ascertain whether they shared similar views or if they differed; any difference may be an important factor in understanding the washback of rSECEE on teaching and learning.

## **5. COLT**

Communicative Orientation of Language Teaching Observation Scheme (COLT) (Spada & Froehlich, 1995). This tool was initially developed by a group of Canadian researchers who investigated the extent to which language classrooms had employed communicative language teaching features. Through employing COLT Part A the researcher was able to thoroughly note all the lesson's activities and episodes including the proportion of time spent on each activity or episode. The use of COLT Part A provided a macroscopic description (Hayes & Read, 2004) of the EFL Libyan classrooms, in relation to the effect of revised SECEE on teaching and learning.

Appendix R: An Example of an Observational Coding Sheet

Observational scheme *Gr 6-8, 1999*

| Time | Activity | Talking | materials used | Comments   |
|------|----------|---------|----------------|--|
| 9:8  | Grammar  | WTS     | textbook       | She revised the title of Reported speech, and she asked the students to give an example  |
| 9:15 | Grammar  | WTS     | textbook       | She focus more on examples - She asked the students to write the example on the board.<br>(3) - interpretation -<br>(3) - interpretation -<br>The teacher write the rest of the examples on the board. |
| 9:25 | "        | IW      |                | - She wrote an example for students to answered by themselves.<br>(3) (3)<br>* she asked them to answer an example as homework   |

(3) Goes through textbook exercise

I want to go on holiday but I can't afford it" Mary said.

## Appendix S: Samples of Teachers' Participants classroom assessment

Teacher A

Teacher B

### SECOND TERM ENGLISH EXAM THIRD YEAR

#### SCIENTIFIC SECTION

DATE :23th/4/2017 ( DON'T WORRY .....FEEL RELAX)

Name.....Class.....

#### Q1- Put (T) for true and (F) for false:(14Marks)

- a-I asked him if when he was born.(        )
- b-Jules Verne was born in Italy in 1818.(        )
- c-We use the infinitive form after a noun or a pronoun to explain the purpose of (sth). (        )
- d-Our teacher is very strict , He expects us to obey him immediately.(        )
- e-The common date palm , phoenix dactylifera is found in north Africa only.(        )
- f-The king of Sicily wanted Aldrisi to produce a map of the world.(        )
- g-I enjoy read English books.(        )
- h-A reported question has the form of a question.(        )
- i-Ibn Sina translated his book the Canon of medicine into Latin in 15<sup>th</sup> century . (        )
- k-The plant kingdom was divided into three phyla.(        )
- l-Albert Einstein suggested the idea of stimulates emission of laser.(        )
- m-The book Principia is written by Alkharizmi. (        )
- n-Because I liked sports , so I joined the club.(        )
- o-Dictoyledon has broad leaves with a main vein .(        )

#### Q2-Choose the best answer:(14 Marks)

- 1-By.....his computer, I could finish quickly. ( to use / use / using)
- 2-The shadow clock was used for measuring time at ( day / night / midnight)
- 3-They enjoy.....TV .( watching / to watch / watch)

## Samples of Teachers' Participants classroom assessment

**Teacher C**

### *The Final Examination Of The First Period for Literary Section*

**Student's Name:**.....

**Class:**

.....

\*\*\*\*\*  
\*\*\*\*\*

#### **Q1: Write (T) for True and (F) for False sentences. Correct the false sentences:**

- 1/ " **Thunder storm**" means light sporadic rain. [   ]
- 2/ The Nazca Lines can't be seen only from a plane high above the ground. [   ]
- 3/ Some of the Nazca Lines are pictures of animals and birds. [   ]
- 4/ " **T-Rex**" is a kind of dinosaurs. [   ]
- 5/ The **Colosseum** was built by Egyptian. [   ]
- 6/ Bent Pyramid is the first attempt of building the pyramids. [   ]
- 7/ "**Tobacco**" is a nasty white stuff. [   ]
- 8/ "**It's a good idea to .....**" is used for giving advice. [   ]
- 9/ Geological time is divided into three eras. [   ]
- 10/ Dinosaurs laid eggs. [   ]
- 11/ All creatures on Earth extinct at the same time as the dinosaurs did. [   ]
- 12/ Fish have existed through all three eras. [   ]
- 13/ The dinosaurs can't have disappeared so suddenly, they disappeared gradually. [   ]
- 14/ **Monsieur Champollion's** task was very difficult. [   ]
- 15/ The use of metal is an indication of progress. [   ]
- 16/ **Bronze is** a mixture of gold and copper. [   ]
- 17/ Successful animals are animals that have been able to survive. [   ]
- 18/ It was thought that birds evolved from dinosaurs over millions of years. [   ]
- 19/ Some dinosaurs were plant eaters and some ate meat. [   ]
- 20/ Dinosaurs were reptiles, [   ]

5

**MARKS**

\*\*\*\*\*  
\*\*\*\*\*

#### **Q2: Choose the correct answer from the given options:**

- 21/ Nisreen will be good ..... Math.      **a-** in.      **b-** at.      **c-** on.
- 22/ Asma ..... to University at this tomorrow.      **a-** will drive.      **b-** – will be driven.      **c-** will driven.
- 23/ ..... places near the sea.      **a-** shower.      **b-** coastal areas.      **c-** flash flood.
- 24/ He is afraid .....dogs.      **a-** of.      **b-** from.      **c-** for.
- 25/ I watched a ..... .      **a-** nice, old and American film.      **b-** American, nice and old film.      **c-** film, nice, old.
- 26/ The lake is ..... clod to swim in.      **a-** so.      **b-** enough.      **c-** too.
- 27/ You were there, .....?      **a-** were you.      **b-** weren't you.      **c-** aren't you.
- 28/ A snake..... the marks on the sand.      **a-** must make.      **b-** must have made.      **c-** must made.
- 29/ Tomorrow.....      **a-** January.      **b-** morning.      **c-** next week.
- 30/ Don't call me at 2:00, because I will be ..... .      **a-** having.      **b-** have.      **c-** haven.
- 31/ The..... theory is much more likely to be true.      **a-** Luis and Walters'.      **b-** Poinars'.      **c-** Peter's Schulte.
- 32/ Dinosaurs are ..... .      **a-** cold-blooded.      **b-** hot-blooded.      **c-** no-blooded.
- 33/ Brooches, necklaces, rings are ..... .      **a-** food.      **b-** jewellery.      **c-** weapons.
- 34/ ..... water.      **a-** vapour.      **b-** fresh.      **c-** drinking.
- 35/ ..... climate.      **a-** tropical.      **b-** summer.      **c-** cold.
- 36/ Forest ..... .      **a-** fire.      **b-** hot.      **c-** cold.
- 37/ Insect ..... .      **a-** groups.      **b-** swarms.      **c-** folks.
- 38/ " Within seconds" means ..... .      **a-** quickly.      **b-** slowly.      **c-** angry.
- 39/ "Terrible" means ..... .      **a-** horrible.      **b-** beautiful.      **c-** terrible.
- 40/ The stone must ..... transported by boat.      **a-** have.      **b-** have been.      **c-** have be.
- 41/ "Tamed" means ..... .      **a-** wild.      **b-** domesticated.      **c-** trained.

42/ The fertile Crescent between the rivers Tigris and Euphrates was known for its..... . a- fishing. b- farming.  
c- singing.

43/ ..... is used to make clothes, fabric, paper, building. a- oil. b- iron.  
c- plants.

44/ Peking man, homo sapiens, homo hobilis, modern man are . origins of..... . a- plant. b- birds  
c- man.

6

**MARKS**

\*\*\*\*\*  
\*\*\*\*\*

**Q3: Match the words:**

|   |  |
|---|--|
| 45/ Becomes water means .....                                 | a- extinct.                                |
| 46/ fish, reptiles, birds, mammals are .....                  | b- vertebrates.                            |
| 47/ If a creature doesn't exist anymore, are called .....     | c- temporary conditions in the atmosphere. |
| 48/ ceremonies means .....                                    | d- water.                                  |
| 49/ Weather means .....                                       | e- adjectives.                             |
| 50/ The scientific formula for water is .....                 | f- melts.                                  |
| 51/ 70% of the human body weight is .....                     | g- adverbs of certainty and uncertainty.   |
| 52/ Remains of animals that are found in rocks called...      | h- dust storms.                            |
| 53/ ( Rainy, cold, hot, snowy, sunny, foggy) are .....        | i- study the past.                         |
| 54/ Archaeologists .....                                      | j- formal social occasion.                 |
| 55/ ( might, clearly, possibly, definitely, likely) are ..... | k- fossils.                                |
| 56/ The ghibli can create .....                               | l- H <sub>2</sub> O.                       |

**3 MARKS**

\*\*\*\*\*  
\*\*\*\*\*

**Q4: Rewrite the sentences:**

57/ an/ He/ is/ intelligent/ clearly/ boy.  
.....

58/ your/ is/ name/ What?  
.....

59/ she/ is/ Who?  
.....

60/ / have/ probably/ They/ got/ cars.  
..... **2 MARKS**

Appendix T: Samples of Teachers C's white board work

